

# Generalized Multiple Importance Sampling

Víctor Elvira<sup>\*</sup>, Luca Martino<sup>†</sup>, David Luengo<sup>‡</sup>, Mónica F. Bugallo<sup>§</sup>

## Abstract

Importance Sampling methods are broadly used to approximate posterior distributions or some of their moments. In its standard approach, samples are drawn from a single proposal distribution and weighted properly. However, since the performance depends on the mismatch between the targeted and the proposal distributions, several proposal densities are often employed for the generation of samples. Under this Multiple Importance Sampling (MIS) scenario, many works have addressed the selection or adaptation of the proposal distributions, interpreting the sampling and the weighting steps in different ways. In this paper, we establish a general framework for sampling and weighing procedures when more than one proposal are available. The most relevant MIS schemes in the literature are encompassed within the new framework, and, moreover novel valid schemes appear naturally. All the MIS schemes are compared and ranked in terms of the variance of the associated estimators. Finally, we provide illustrative examples which reveal that, even with a good choice of the proposal densities, a careful interpretation of the sampling and weighting procedures can make a significant difference in the performance of the method.

## Index Terms

Monte Carlo Methods, Multiple Importance Sampling, Bayesian Inference.

## I. INTRODUCTION

Importance Sampling (IS) is a well-known Monte Carlo technique applied to compute integrals involving target probability density functions (pdfs) [1, 2]. The standard IS technique draws samples from a single proposal pdf and assigns them weights based on the ratio between the target and the proposal pdfs, both evaluated at the sample value. The choice of a suitable proposal pdf is crucial for obtaining a good approximation of the target pdf using the IS method. Indeed, although the validity of this approach is guaranteed under mild assumptions, the variance of the estimator depends notably on the discrepancy between the shape of the proposal and the target [1, 2].

Several advanced strategies have been proposed in the literature for designing efficient IS schemes [2, Chapter 2], [3, Chapter 9], [4]. A powerful approach is based on using a population of different proposal pdfs. This approach is often referred to as *multiple* importance sampling (MIS) and several possible implementations have been proposed [5, 6, 7]. In general, MIS strategies provide more robust algorithms, since they avoid to entrust the performance of the method to a single proposal. Moreover, many algorithms have been proposed in order to conveniently adapt the set of proposals in MIS [8, 9, 10].

When a set of proposal pdfs is available, the way of sampling and weighting the samples is not unique, unlike the case of using a single proposal. Indeed, different MIS algorithms in the literature (both adaptive and not adaptive) have implicitly and independently interpreted the sampling and weighting procedures in different ways [6, 8, 9, 7, 10, 11, 12]. Namely, there are several possible combinations of sampling and weighting schemes when a set of proposal pdfs is available, which lead to valid

<sup>\*</sup>Víctor Elvira is with the Dept. of Signal Theory and Communications, Universidad Carlos III de Madrid (Spain). Email: velvira@tsc.uc3m.es

<sup>†</sup>Luca Martino is with the Institute of Mathematical Sciences and Computing, University of São Paulo (Brazil). Email: lukafree@icmc.usp.br

<sup>‡</sup>David Luengo is with the Dept. of Signal Theory and Communications, Universidad Politécnica de Madrid (Spain). Email: david.luengo@upm.es

<sup>§</sup>Mónica F. Bugallo is with the Dept. of Electrical and Computer Engineering, Stony Brook University (USA) Email: monica.bugallo@stonybrook.edu

MIS approximations of the target pdf. However, these different possibilities can largely differ in terms of performance of the corresponding estimators.

In this paper, we introduce a unified framework for MIS schemes, providing a general theoretical description of the possible sampling and weighting procedures when a set of proposal pdfs is used for producing an IS approximation. Within this unified context, it is possible to interpret that all the MIS algorithms draw samples from a equal-weighted mixture distribution obtained from the set of available proposal pdfs. Three different sampling approaches and five different functions to calculate the weights of the generated samples are proposed and discussed. Moreover, we state two basic rules for possibly devising new valid sampling and weighting strategies within the proposed framework. All the analyzed combinations of sampling/weighting provide consistent estimates of the parameters of interest.

The proposed generalized framework includes all of the existing MIS methodologies that we are aware of (applied within different algorithms, e.g. in [7, 8, 11, 10, 12]) and allows the design of novel techniques (here we propose three new schemes, but more can be introduced). An exhaustive theoretical analysis is provided by introducing general expressions for sampling and weighting in this generalized MIS context, and by proving that they yield consistent estimators. Furthermore, we compare the performance of the different MIS schemes (the proposed and the existing ones) in terms of the variance of the estimators and effective sample size. We illustrate the performance of all the methods by means of three numerical examples, highlighting the differences among the different MIS schemes in terms of performance and computational cost. In particular, in the first two examples, where the proposal pdfs are intentionally well chosen, evidence the significative effects produced by the different interpretations of the sampling and weighting schemes in MIS.

The rest of this paper is organized as follows. In Section II, we describe the problem and we revisit the standard IS methodology. In Section III, we discuss the sampling procedure in MIS, propose three new sampling strategies, and analyze some distributions of interest. In Section IV, we propose five different weighting functions, some of them completely new, and show their validity. The different combinations of sampling/weighting strategies are analyzed in Section V, establishing the connections with existent MIS schemes, and describing three novel MIS schemes. In Section VI, we analyze the performance of the different MIS schemes in terms of the variance of the estimators and the effective sample size. Section VIII presents some descriptive numerical examples where the different MIS schemes are simulated, and finally, Section IX contains some concluding remarks.

## II. PROBLEM STATEMENT AND BACKGROUND

Let us consider a system characterized by a vector of  $d_x$  unknown parameters,  $\mathbf{x} \in \mathbb{R}^{d_x}$ , and a set of  $d_y$  observed data made about the system,  $\mathbf{y} \in \mathbb{R}^{d_y}$ .<sup>1</sup> A general objective is to extract the complete information about the latent state,  $\mathbf{x}$ , given the observations,  $\mathbf{y}$ , by means of studying the posterior density function (pdf) defined as

$$\tilde{\pi}(\mathbf{x}|\mathbf{y}) = \frac{\ell(\mathbf{y}|\mathbf{x})h(\mathbf{x})}{Z(\mathbf{y})} \propto \pi(\mathbf{x}|\mathbf{y}) = \ell(\mathbf{y}|\mathbf{x})h(\mathbf{x}), \quad (1)$$

where  $\ell(\mathbf{y}|\mathbf{x})$  is the likelihood function,  $h(\mathbf{x})$  is the prior pdf, and  $Z(\mathbf{y})$  is the normalization factor.<sup>2</sup> The objective is to approximate the pdf of interest (referred to as target pdf) by Monte Carlo-based sampling [1, 2, 3]. The resulting approximation of  $\pi(\mathbf{x})$  will be denoted as  $\hat{\pi}(\mathbf{x})$  and will be attained using importance sampling (IS) techniques.

<sup>1</sup>Vectors are denoted by bold-faced letters, e.g.,  $\mathbf{x}$ , while regular-faced letters are used for scalars, e.g.,  $x$ .

<sup>2</sup>In the sequel, to simplify the notation, the dependence on  $\mathbf{y}$  is removed, e.g.,  $Z \equiv Z(\mathbf{y})$ .

### A. Standard Importance Sampling

Importance Sampling is a general Monte Carlo technique for the approximation of a pdf of interest by a random measure composed of samples and weights [1]. In its original formulation, a set of  $N$  samples,  $\{\mathbf{x}_n\}_{n=1}^N$ , is drawn from a single proposal pdf,  $q(\mathbf{x})$ , characterized by tails that are heavier than those of the target pdf,  $\pi(\mathbf{x})$ . A particular sample,  $\mathbf{x}_n$ , is assigned a weight,  $w_n$ , which measures the adequacy of that particular sample in the approximation of the posterior pdf. Namely, the importance weight is given by

$$w_n = \frac{\pi(\mathbf{x}_n)}{q(\mathbf{x}_n)}, \quad n = 1, \dots, N, \quad (2)$$

which represents the ratio between the target pdf,  $\pi$ , and the proposal pdf,  $q$ , both evaluated at  $\mathbf{x}_n$ . The samples and the weights form the random measure  $\chi = \{\mathbf{x}_n, w_n\}_{n=1}^N$  that approximates the measure of the target pdf as

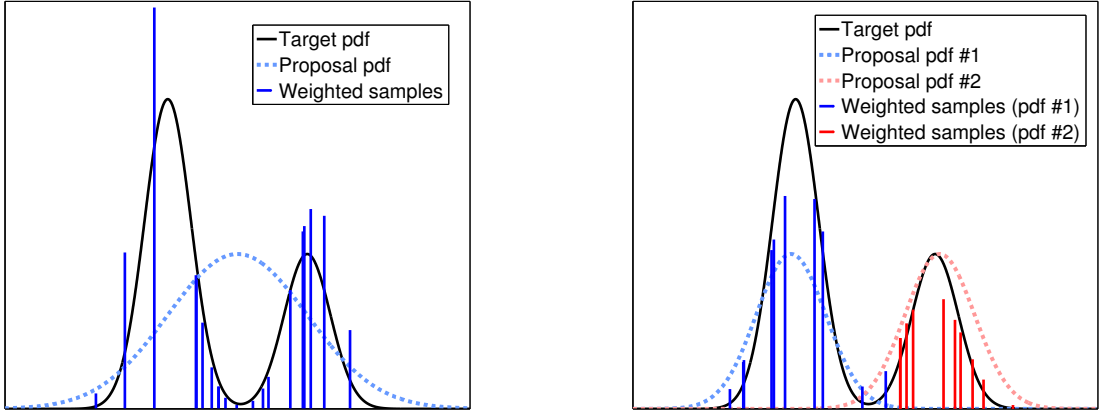
$$\hat{\pi}_{\text{IS}}(\mathbf{x}) = \frac{1}{N\hat{Z}} \sum_{n=1}^N w_n \delta_{\mathbf{x}_n}(\mathbf{x}), \quad (3)$$

where  $\delta_{\mathbf{x}_n}(\mathbf{x})$  is the unit delta measure concentrated at  $\mathbf{x}_n$  and  $\hat{Z} = \frac{1}{N} \sum_{j=1}^N w_j$  is an unbiased estimator of  $Z = \int \pi(\mathbf{x}) d\mathbf{x}$  [1]. Fig. 1 (a) displays an example of a target pdf and a proposal pdf, as well as the samples and weights that form a random measure approximating the posterior.

Although the weights of Eq. (2) are broadly used in the literature, the concept of a *properly weighted sample*, suggested in [1, Section 14.2] and in [2, Section 2.5.4], can be used to construct more general weights. More specifically, following the definition in [2, Section 2.5.4], a set of weighted samples is considered *proper* with respect to the target  $\pi$  if, for any square integrable function  $g$ ,

$$E_q[w_n g(\mathbf{x}_n)] = c E_\pi[g(\mathbf{x})], \quad \forall n = \{1, \dots, N\}, \quad (4)$$

where  $c$  is a constant value, also independent from the index  $n$ , and the expectation of the left hand side is performed w.r.t. to the joint pdf of  $w_n$  and  $\mathbf{x}_n$ ,  $q(w, \mathbf{x})$ . Note that in this way, not only  $\mathbf{x}_n$  but also  $w_n$  (for a given value of  $\mathbf{x}_n$ ) can be a r.v. Thus, one can design any joint  $q(w, \mathbf{x})$  as long as the restriction of Eq. (4) is fulfilled [4]. In the sequel, we extend the concept of properly weighted sample within the context of the MIS approach (see in particular Section IV-A).



(a) Single proposal pdf (standard IS).

(b) Two proposal pdfs (MIS).

Fig. 1: Approximation of the target pdf,  $\pi(\mathbf{x})$ , by the random measure  $\chi$ .

### III. SAMPLING IN MULTIPLE IMPORTANCE SAMPLING

MIS-based schemes consider a set of  $N$  proposal pdfs,<sup>3</sup>

$$\{q_N(\mathbf{x})\}_{n=1}^N \equiv \{q_1(\mathbf{x}), \dots, q_N(\mathbf{x})\}.$$

and proceed by generating  $M$  samples,  $\{\mathbf{x}_m\}_{m=1}^M$  (where  $M \neq N$ , in general) from the set of proposal pdfs and by properly weighting the drawn samples. As a visual example, Fig. 1 (b) displays a target pdf and two proposal pdfs as well as the samples and weights that form a random measure approximating the posterior.

It is in the way that the sampling and the weighting are performed that different variants of MIS can be devised. In this section we focus on the generation of samples  $\{\mathbf{x}_m\}_{m=1}^M$ . For clarity in the explanations and the theoretical proofs, we consider  $M = N$ , i.e., the number of samples to be generated coincides with the number of proposal pdfs. All the considerations can be automatically extended for the case  $M = kN$ , with  $k \geq 1$  and  $k \in \mathbb{N}$ .

Note that the use of the complete set of  $N$  proposal pdfs can be interpreted as a single proposal equal-weighted mixture of pdfs, i.e.,

$$\psi(\mathbf{x}) \equiv \frac{1}{N} \sum_{n=1}^N q_n(\mathbf{x}). \quad (5)$$

This is an important interpretation for motivating some of the sampling and weighting schemes discussed in this paper. More precisely, here we describe some basic ways of sampling from a mixture of proposal pdfs and we focus on different statistical properties that will be very useful in the different MIS interpretations.

#### A. Sampling from a mixture of proposal pdfs

In order to provide a better explanation of the discussed sampling procedures, we employ a simile with the urn sampling problem. Let us consider an urn that contains  $N$  balls, where each ball is assigned an index  $j \in \{1, \dots, N\}$ , representing the  $j$ -th proposal of the complete set of available proposal pdfs,  $\{q_j(\mathbf{x})\}_{j=1}^N$ . Then, a generic sampling scheme for generating  $N$  samples from  $\psi(\mathbf{x})$  is given below. Starting with  $n = 1$ :

- 1) Draw a ball from the urn, i.e., choose an index  $j_n \in \{1, \dots, N\}$  using some suitable approach. This corresponds to the selection of a proposal pdf,  $q_{j_n}$ .
- 2) Generate a sample  $\mathbf{x}_n$  from the selected proposal pdf, i.e.,  $\mathbf{x}_n \sim q_{j_n}(\mathbf{x}_n)$ .
- 3) Set  $n = n + 1$  and go to step 1.

Therefore, obtaining the set of samples  $\{\mathbf{x}_n\}_{n=1}^N \equiv \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$  from the mixture pdf  $\psi$  is in general a two step sequential procedure. First, the  $n$ -th index  $j_n$  is drawn according to some conditional pdf,  $P(j_n|j_{1:n-1})$ , where  $j_{1:n-1} \equiv \{j_1, \dots, j_{n-1}\}$  is the sequence of the previously generated indexes.<sup>4</sup> Then, the  $n$ -th sample is drawn from the selected proposal pdf as  $\mathbf{x}_n \sim p(\mathbf{x}_n|j_n)$ . Within this context, one can formulate the joint probability distribution of the current sample and all indexes used to generate the samples from 1 to  $n$  as

$$\begin{aligned} p(\mathbf{x}_n, j_{1:n}) &= P(j_{1:n-1})P(j_n|j_{1:n-1})p(\mathbf{x}_n|j_n) \\ &= P(j_1) \left[ \prod_{i=2}^n P(j_i|j_{1:i-1}) \right] p(\mathbf{x}_n|j_n), \end{aligned} \quad (6)$$

<sup>3</sup>The term normalized refers to the fact that the integral of each of the individual proposal pdfs equals to one.

<sup>4</sup>We use simplified argument-wise notation where  $p(\mathbf{x}_n)$  denotes the pdf of the continuous random variable (r.v.)  $\mathbf{X}_n$ , while  $P(j_n)$  denotes the probability mass function (pmf) of the discrete r.v.  $J_n$ . Also,  $p(\mathbf{x}_n, j_n)$  denotes the joint pdf and  $p(\mathbf{x}_n|j_n)$  is the conditional pdf of  $\mathbf{X}_n$  given  $J_n = j_n$ . If the argument of  $p(\cdot)$  is different from  $\mathbf{x}_n$ , then it denotes the evaluation of the pdf as a function, e.g.,  $p(\mathbf{z}|j_n)$  denotes the pdf  $p(\mathbf{x}_n|j_n)$  evaluated in  $\mathbf{z}$ .

where

$$p(\mathbf{x}_n | j_n) = q_{j_n}(\mathbf{x}_n), \quad (7)$$

for all  $n \in \{1, \dots, N\}$  represents the conditional pdf of the  $n$ -th sample given the  $n$ -th selected index, i.e., the selected proposal pdf,  $q_{j_n}(\mathbf{x}_n)$ . The full joint distribution of all samples and indexes is given by

$$p(\mathbf{x}_{1:N}, j_{1:N}) = P(j_1) \left[ \prod_{i=2}^N P(j_i | j_{1:i-1}) \right] \left[ \prod_{i=1}^N p(\mathbf{x}_i | j_i) \right]. \quad (8)$$

The corresponding graphical model is depicted in Fig. 2.

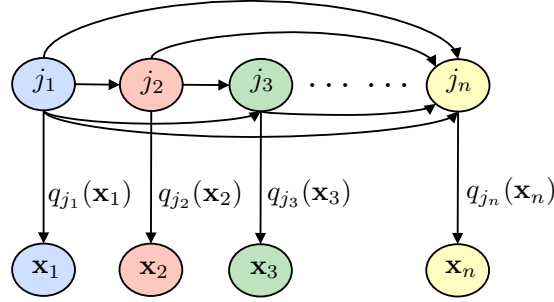


Fig. 2: Graphical model associated to the generic sampling scheme.

### B. Selection of the proposal pdfs

In the sequel, we describe three mechanisms for obtaining the sequence of indexes,  $j_{1:N} \equiv j_1, \dots, j_N$  of the proposal pdfs: two random mechanisms (with and without replacement) and a deterministic scheme which is a particular case of special interest. The resulting sampling methods will be labeled as  $\mathcal{S}_1$ ,  $\mathcal{S}_2$ , and  $\mathcal{S}_3$ , respectively, for easier reference later in the paper. All the mechanisms share the property that

$$\frac{1}{N} \sum_{n=1}^N P(J_n = k) = \frac{1}{N}, \quad \forall k \in \{1, \dots, N\}, \quad (9)$$

i.e., all the indexes have the same probability of being selected.

$\mathcal{S}_1$ : *Random index selection with replacement*:

This is the standard sampling scheme where  $N$  indexes are independently drawn from the set  $\{1, \dots, N\}$ , i.e., from the multinomial distribution defined by the  $N$  possible values, each with probability  $1/N$ . Thus, in this case, we have

$$P(j_n | j_{1:n-1}) = P(j_n) = \frac{1}{N}. \quad (10)$$

With this type of index sampling, several  $j_n$ 's may take the same value, i.e., there may ultimately be more than one sample,  $\mathbf{x}_n$  with  $n = 1, \dots, N$ , generated from the same proposal pdf and there may be proposal pdfs that are not used for generating any samples.

$\mathcal{S}_2$ : *Random index selection without replacement*:

In this case, when an index is selected from the set of available values, that particular index is discarded for future generations of indexes. This can be formulated as the indexes being uniformly and sequentially drawn from different sets,

i.e.,

$$\begin{aligned} j_1 &\in \mathcal{I}_1 = \{1, \dots, N\}, \\ j_2 &\in \mathcal{I}_2 = \{1, \dots, N\} \setminus \{j_1\}, \\ j_3 &\in \mathcal{I}_3 = \{1, \dots, N\} \setminus \{j_1, j_2\}, \end{aligned}$$

and, more generally,  $j_n \in \mathcal{I}_n = \{1, \dots, N\} \setminus \{j_{1:n-1}\}$ . The conditional probability mass function (pmf) of the  $n$ -th index given the previous ones is given by

$$P(J_n = k | j_{1:n-1}) = \begin{cases} \frac{1}{N - n + 1} & \text{if } k \in \mathcal{I}_n, \\ 0 & \text{if } k \notin \mathcal{I}_n, \end{cases} \quad (11)$$

where  $|\mathcal{I}_n| = N - n + 1$ . Note that, the marginal pmf of the  $j$ -th index is again<sup>5</sup>

$$P(j_n) = \frac{1}{N}. \quad (12)$$

Under this scheme, once an index is selected and a sample is generated from the corresponding proposal pdf, no more samples will be generated from that particular proposal pdf. One can observe that at each generation step (i.e., for each sample to be generated), the number of available proposal pdfs is reduced (the proposal pdfs that were previously used are discarded). Note that with this strategy, exactly one sample is drawn from each of the possible proposal pdfs.

$\mathcal{S}_3$ : *Deterministic index selection without replacement*: This sampling is a particular case of sampling  $\mathcal{S}_2$  where a fixed deterministic sequence of indexes is drawn. For instance, and without loss of generality,

$$j_1 = 1, j_2 = 2, \dots, j_n = n, \dots, j_N = N.$$

Therefore, the  $n$ -th sample is deterministically drawn from  $q_n(\mathbf{x}_n)$ , i.e.,

$$\mathbf{x}_n \sim q_{j_n}(\mathbf{x}_n) = q_n(\mathbf{x}_n). \quad (13)$$

We can express the conditional pmf of the  $n$ -th index given the  $n - 1$  previous ones as

$$P(j_n | j_{1:n-1}) = P(j_n) = \mathbb{1}_{j_n=n}, \quad (14)$$

where  $\mathbb{1}$  denotes the indicator function. Again, each of the  $N$  proposal pdfs is used to generate one, and only one, sample of the set of samples  $\{\mathbf{x}_n\}_{n=1}^N$ . This particular index selection procedure has been used by different algorithms in the MIS literature (e.g., APIS [10]), and it is also implicitly present in the particle filtering literature (e.g. bootstrap PF [13]).

### C. Distributions of interest of the $n$ -th sample, $\mathbf{x}_n$

In the following, we discuss some important distributions related to the set of generated samples. These distributions are of utmost importance for understanding the different methods for weighting the samples that are discussed in the next section.

The considered sampling framework operates in a sequential way, i.e., samples are generated one after another. Under that perspective, the distribution of the  $n$ -th sample given all the knowledge of the process up to that point is  $p(\mathbf{x}_n | j_{1:n-1}, \mathbf{x}_{1:n-1}) = p(\mathbf{x}_n | j_{1:n-1})$ . This is of particular interest, since this is the pdf of the r.v.  $\mathbf{X}_n$  after the  $n - 1$  precedent samples have been drawn (since the sequence  $j_{1:n-1}$  is known). In the random index selection with replacement ( $\mathcal{S}_1$ ), this distribution corresponds to

<sup>5</sup>There are  $N!$  equiprobable configurations (permutations) of the sequence  $\{j_1, \dots, j_N\}$ , and in  $(N - 1)!$  the  $k$ -th index is drawn at the  $n$ -th position,  $\forall k, n = 1, \dots, N$ . Therefore  $P(J_n = k) = \frac{(N-1)!}{N!} = \frac{1}{N} \forall k, n = 1, \dots, N$ .

$p(\mathbf{x}_n | j_{1:n-1}) = \psi(\mathbf{x}_n)$ . For the random index selection without replacement ( $\mathcal{S}_2$ ), we have  $p(\mathbf{x}_n | j_{1:n-1}) = \frac{1}{|\mathcal{I}_n|} \sum_{k \in \mathcal{I}_n} q_k(\mathbf{x})$ . Finally, under the deterministic index selection scheme ( $\mathcal{S}_3$ ),  $p(\mathbf{x}_n | j_{1:n-1}) = q_n(\mathbf{x}_n)$ .

Once the  $n$ -th index  $j_n$  has been selected, the  $n$ -th sample,  $\mathbf{x}_n$ , is distributed as  $p(\mathbf{x}_n | j_n) = q_{j_n}(\mathbf{x}_n)$ . Note that this is common to any valid sampling method within the proposed framework. The marginal distribution of this  $n$ -th sample  $\mathbf{x}_n$  is, in general and regardless the specific sampling procedure, given by

$$\begin{aligned} p(\mathbf{x}_n) &= \sum_{k=1}^N p(\mathbf{x}_n | J_n = k) P(J_n = k) \\ &= \sum_{k=1}^N q_k(\mathbf{x}_n) P(J_n = k). \end{aligned} \quad (15)$$

While  $p(\mathbf{x}_n | j_n)$  is the same for the three proposed sampling methods (and to any other that one can implement under this framework), the marginal distribution  $P(J_n = k)$  differs for the different methods. Namely, when randomly selecting the indexes (with ( $\mathcal{S}_1$ ) or without ( $\mathcal{S}_2$ ) replacement), this marginal distribution amounts to  $P(J_n = k) = \frac{1}{N}, \forall n, \forall k$ , and therefore  $p(\mathbf{x}_n) = \frac{1}{N} \sum_{k=1}^N q_k(\mathbf{x}_n) = \psi(\mathbf{x}_n)$ . In the case of the deterministic index selection ( $\mathcal{S}_3$ ),  $P(J_n = k) = \mathbb{1}_{k=n}$ , and therefore  $p(\mathbf{x}_n) = q_n(\mathbf{x}_n)$ , i.e., the distribution of the r.v.  $\mathbf{X}_n$  is the  $n$ -th proposal pdf, and not the whole mixture as in the other sampling schemes with random index selection.

#### D. Distributions of interest beyond $\mathbf{x}_n$

In the proposed framework we draw  $N$  samples that we will use, following IS arguments, for approximation of the target distribution or for the calculation of estimators. The traditional IS approach focuses just on the distribution of the r.v.  $\mathbf{X}_n$ . Here, we are also interested in the distribution of the samples regardless of their index  $n$ . The reason is that, in MIS schemes, the  $N$  samples can be used jointly regardless of their specific order of appearance. More precisely, we introduce a generic r.v. defined as

$$\mathbf{X} = \mathbf{X}_n \quad \text{with} \quad n \sim \mathcal{U}\{1, 2, \dots, N\}, \quad (16)$$

where  $\mathcal{U}\{1, 2, \dots, N\}$  is the discrete uniform distribution on the set  $\{1, 2, \dots, N\}$ . Namely, the r.v.  $\mathbf{X}$  is equal to  $\mathbf{X}_n$  chosen uniformly within the set  $\{\mathbf{X}_1, \dots, \mathbf{X}_N\}$ . Therefore, the density of  $\mathbf{X}$  is given by the expression<sup>6</sup>

$$f(\mathbf{x}) = \frac{1}{N} \sum_{n=1}^N p_{\mathbf{x}_n}(\mathbf{x}), \quad (17)$$

where  $p_{\mathbf{x}_n}(\mathbf{x})$  denotes the marginal pdf of  $\mathbf{X}_n$ , given by Eq. (15), evaluated at  $\mathbf{x}$ .<sup>7</sup> In the sampling schemes with random index selection ( $\mathcal{S}_1$  and  $\mathcal{S}_2$ ), since  $p_{\mathbf{x}_n}(\mathbf{x}) = \psi(\mathbf{x})$ , we also have  $f(\mathbf{x}) = \frac{1}{N} \sum_{n=1}^N \psi(\mathbf{x}) = \psi(\mathbf{x})$ , whereas in the sampling with deterministic index selection ( $\mathcal{S}_3$ ), since  $p_{\mathbf{x}_n}(\mathbf{x}) = q_n(\mathbf{x})$ , we have  $f(\mathbf{x}) = \frac{1}{N} \sum_{n=1}^N q_n(\mathbf{x}) = \psi(\mathbf{x})$ .

Note that, in all the sampling schemes,  $\mathbf{X}$  is distributed according to the mixture  $\psi(\mathbf{x})$ , as expected. Moreover, one can also obtain the conditional pdf of  $\mathbf{X}$  given the sequence of indexes as

$$f(\mathbf{x} | j_{1:N}) = \frac{1}{N} \sum_{k=1}^N p_{\mathbf{x}_k}(\mathbf{x} | j_{1:N}) = \frac{1}{N} \sum_{k=1}^N q_{j_k}(\mathbf{x}). \quad (18)$$

Note that in this case,  $f(\mathbf{x} | j_{1:N}) = \psi(\mathbf{x})$  for the schemes without replacement at the index selection ( $\mathcal{S}_2$  and  $\mathcal{S}_3$ ), but  $f(\mathbf{x} | j_{1:N}) = \frac{1}{N} \sum_{n=1}^N q_{j_n}(\mathbf{x})$  for the case with replacement ( $\mathcal{S}_1$ ), i.e., conditioned to the selection of the indexes, some proposal pdfs may not appear while others may appear repeated.

<sup>6</sup>Again, we use a simplified argument-wise notation where  $f(\mathbf{x})$  denotes the pdf of the r.v.  $\mathbf{X}$  of Eq. (16).

<sup>7</sup>In Eq. (17), we have used the notation  $p_{\mathbf{x}_n}(\mathbf{x})$ , instead of  $p(\mathbf{x})$  as in Eq. (15), for denoting the marginal pdf of  $\mathbf{X}_n$  evaluated in  $\mathbf{x}$ . However, in the rest of the paper, we use the simpler one,  $p(\mathbf{x}_n)$ .

TABLE I: Summary of the distributions of the r.v.'s  $J_n$ ,  $\mathbf{X}_n$  and  $\mathbf{X}$ , for the three different sampling procedures.

|   | Selection of the indexes                       |   |   |                       |
|---|--|---|---|-----------------------|
| Distributions   | With replacement<br>$\mathcal{S}_1$            | Without Replacement   |   | Text references       |
|   |  | random selection<br>$\mathcal{S}_2$   | deterministic selection<br>$\mathcal{S}_3$  |                       |
|   |  |   |   |                       |
| $J_n \sim P(j_n)$   | $\frac{1}{N}$                                  | $\frac{1}{N}$   | $\mathbb{1}_{j_n=n}$                        | Eqs. (10)-(12)-(14)   |
| $J_n   J_{1:n-1} \sim P(j_n   j_{1:n-1})$                   | $\frac{1}{N}$                                  | $\frac{1}{ \mathcal{I}_n } \mathbb{1}_{j_n \in \mathcal{I}_n}$  | $\mathbb{1}_{j_n=n}$                        | Eqs. (10)-(11)-(14)   |
| $\mathbf{X}_n   J_{1:n-1} \sim p(\mathbf{x}_n   j_{1:n-1})$ | $\psi(\mathbf{x}_n)$                           | $\frac{1}{ \mathcal{I}_n } \sum \forall k \in \mathcal{I}_n q_k(\mathbf{x}_n)$                                  | $q_n(\mathbf{x}_n)$                         | Sect. III-C           |
| $\mathbf{X}_n   J_n \sim p(\mathbf{x}_n   j_n)$             | $q_{j_n}(\mathbf{x}_n)$                        | $q_{j_n}(\mathbf{x}_n)$   | $q_{j_n}(\mathbf{x}_n) = q_n(\mathbf{x}_n)$ | Eq. (7)               |
| $\mathbf{X}_n \sim p(\mathbf{x}_n)$                         | $\psi(\mathbf{x}_n)$                           | $\psi(\mathbf{x}_n)$  | $q_n(\mathbf{x}_n)$                         | Eq. (15)              |
| $\mathbf{X}   J_{1:N} \sim f(\mathbf{x}   j_{1:N})$         | $\frac{1}{N} \sum_{n=1}^N q_{j_n}(\mathbf{x})$ | $\psi(\mathbf{x})$  | $\psi(\mathbf{x})$                          | Eq. (18)              |
| $\mathbf{X} \sim f(\mathbf{x})$                             | $\psi(\mathbf{x})$                             | $\psi(\mathbf{x})$  | $\psi(\mathbf{x})$                          | Eq. (17)              |
| $\mathbf{X}_{1:N} \sim p(\mathbf{x}_{1:N})$                 | $\prod_{n=1}^N \psi(\mathbf{x}_n)$             | $\psi(\mathbf{x}_1) \prod_{n=2}^N \frac{1}{ \mathcal{I}_n } \sum_{\ell \in \mathcal{I}_n} q_\ell(\mathbf{x}_n)$ | $\prod_{n=1}^N q_n(\mathbf{x}_n)$           | Sect. III-D; Eq. (19) |

**Remark 1.** (Sampling): In the proposed framework, we consider valid, any sequential sampling scheme for generating the set  $\{\mathbf{X}_1, \dots, \mathbf{X}_N\}$  such that the pdf of the r.v.  $\mathbf{X}$  defined in Eq. (16) is given by  $\psi(\mathbf{x})$ . Further considerations about the r.v.  $\mathbf{X}$  and connections with variance reduction methods [1, 3] are given in Appendix A.

Table I summarizes the discussed distributions of interest related to the different index selection schemes. In particular, different distributions of the r.v.'s indexes  $J_n$ , and samples  $\mathbf{X}_n$  and  $\mathbf{X}$  are shown. Note that, related to the previous remark, in the three considered index selection procedures presented in this section, the pdf of the r.v.  $\mathbf{X}$  is the mixture  $\psi(\mathbf{x})$ . However, different procedures yield different conditional and marginal distributions that will be exploited to justify different strategies for calculation of the importance weights in the next section. Finally, the last row of the table shows the joint distribution  $p(\mathbf{x}_{1:N})$  of the variables  $\mathbf{X}_1, \dots, \mathbf{X}_N$ , i.e.,  $p(\mathbf{x}_{1:N}) = \prod_{n=1}^N \psi(\mathbf{x}_n)$  and  $p(\mathbf{x}_{1:N}) = \prod_{n=1}^N q_n(\mathbf{x}_n)$  for the sampling with replacement ( $\mathcal{S}_1$ ) and deterministic selection ( $\mathcal{S}_3$ ), respectively. For the sampling without replacement and random selection this joint distribution is

$$p(\mathbf{x}_{1:N}) = \psi(\mathbf{x}_1) \prod_{n=2}^N \frac{1}{|\mathcal{I}_n|} \sum_{\ell \in \mathcal{I}_n} q_\ell(\mathbf{x}_n), \quad (19)$$

with  $\mathcal{I}_n = \{1, \dots, N\} \setminus \{j_{1:n-1}\}$ .

#### IV. WEIGHTING IN MIS

The weighting step is used to evaluate the adequacy of the samples generated from the proposal pdfs with respect to the general objective of approximating the target  $\pi(\mathbf{x})$ . The weight assigned to the  $n$ -th sample is proportional to the ratio between the target pdf evaluated at the sample value,  $\pi(\mathbf{x}_n)$ , and the proposal pdf evaluated at the sample value, i.e.,

$$w_n = \frac{\pi(\mathbf{x}_n)}{\varphi_{\mathcal{P}_n}(\mathbf{x}_n)}, \quad (20)$$

where the generic function  $\varphi_{\mathcal{P}_n}$  represents the proposal pdf from which it is interpreted that the  $n$ -th sample is drawn. In general, this function may be parameterized by a subset or the entire sequence of indexes  $j_{1:N}$ , i.e.,  $\mathcal{P}_n \subseteq \{j_1, \dots, j_N\}$  (further details are given below).

It is on this interpretation of what the proposal pdf used for the generation of the sample is (the evaluation of the denominator in the weight calculation) that different weighting strategies can be devised.



### A. Mathematical justification

Let us consider the integral  $I = \int g(\mathbf{x})\pi(\mathbf{x})d\mathbf{x}$ , where  $\pi(\mathbf{x})$  is the target distribution and  $g$  is any square integrable function w.r.t.  $\pi(\mathbf{x})$ . The generic IS estimator of  $I$  is given by

$$\hat{I} = \frac{1}{NZ} \sum_{n=1}^N w_n g(\mathbf{x}_n) \quad (21)$$

where  $w_n$  is the importance weight of the  $n$ -th sample,  $\mathbf{x}_n$ , and  $Z = \int \pi(\mathbf{x})d\mathbf{x}$  is the normalizing constant. In standard IS, when  $Z$  is known,  $\hat{I}$  is an unbiased estimator of  $I$ . Otherwise, if the target distribution is only known up to the normalizing constant,  $Z$ , then one can estimate it as

$$\hat{Z} = \frac{1}{N} \sum_{n=1}^N w_n, \quad (22)$$

which is an unbiased estimator and asymptotically consistent with the number of samples under some mild assumptions regarding the tails of the proposal and target distributions [1]. Therefore,  $\hat{I}$  is also asymptotically consistent, even when  $Z$  is unknown and is replaced with  $\hat{Z}$  instead [1].

We recall that in standard IS, there is just one proposal pdf,  $q(\mathbf{x})$ , from which the samples are drawn, and therefore, the weight is always calculated as  $w_n = \frac{\pi(\mathbf{x}_n)}{q(\mathbf{x}_n)}$  for  $n = 1, \dots, N$ . The extension to MIS is not straightforward: for instance, one could reasonably state that  $\mathbf{x}_n$  comes from the marginal distribution  $p(\mathbf{x}_n)$  (which depends on the specific sampling scheme), while  $p(\mathbf{x}_n | j_n) = q_{j_n}(\mathbf{x}_n)$  could also be intuitively considered as the proposal distribution if the realization  $j_n$  of the r.v.  $J_n$  is observed (as it always happens).

Our approach is based on analyzing which weighting functions yield *proper* estimators. We consider the definition of properness by Liu [2, Section 2.5] and we extend (or relax) it to the MIS scenario. Namely, Liu-properness in standard IS states that a weighted sample  $\{\mathbf{x}_n, w_n\}$  drawn from a single proposal  $q$  is proper if, for any square integrable function  $g$ ,

$$\frac{E_q[g(\mathbf{x})w(\mathbf{x})]}{E_q[\pi(\mathbf{x})]} = E_\pi[g(\mathbf{x})], \quad (23)$$

i.e.,  $w$  can be in any form as long as the condition of Eq. (23) is fulfilled. Note that, for a deterministic weight assignment, the only proper weights are the ones considered by the standard IS approach. However, many different proper weights can be designed if a probabilistic weight function is considered for a given value of  $\mathbf{x}$ .

The extension to the MIS scenario is not straightforward, since there are several valid sampling procedures and, in each of them, different interpretations of the proposal pdf of each sample. Therefore, we propose a generalized properness condition in the MIS scenario over the whole estimator. Namely, given a specific sampling method, we consider that the set of weighting functions  $\{w_n\}_{n=1}^N$  is proper if

$$\frac{E_{p(\mathbf{x}_{1:N}, j_{1:N})} \left[ \frac{1}{N} \sum_{n=1}^N w_n g(\mathbf{x}_n) \right]}{E_{p(\mathbf{x}_{1:N}, j_{1:N})} \left[ \frac{1}{N} \sum_{n=1}^N w_n \right]} = E_\pi[g(\mathbf{x})]. \quad (24)$$

This is equivalent to impose the restriction

$$\frac{E_{p(\mathbf{x}_{1:N}, j_{1:N})} [Z\hat{I}]}{E_{p(\mathbf{x}_{1:N}, j_{1:N})} [\hat{Z}]} = I, \quad (25)$$

which is fulfilled if  $E[\hat{I}] = I$  and  $E[\hat{Z}] = Z$ . Note that the MIS properness is fulfilled by any set of weighting functions  $\{w_n\}_{n=1}^N$  that yield an unbiased generic estimator  $\hat{I}$ , i.e.,  $E[\hat{I}] = I$ . Note that the MIS properness is a relaxation of the one proposed by Liu, i.e., any Liu-proper weighting scheme is also proper according to our definition, but not vice versa.

Following the general sampling framework defined in Section III (see Eq. (8)),

$$\begin{aligned} E[\hat{I}] &= \frac{1}{Z} \sum_{j_{1:N}} \int \left( \frac{1}{N} \sum_{n=1}^N w_n g(\mathbf{x}_n) \right) p(\mathbf{x}_{1:N}, j_{1:N}) d\mathbf{x}_{1:N} \\ &= \frac{1}{ZN} \sum_{n=1}^N \sum_{j_{1:N}} \int w_n g(\mathbf{x}_n) P(j_{1:N}) \cdot \left[ \prod_{i=1}^N p(\mathbf{x}_i | j_i) \right] d\mathbf{x}_{1:N}. \end{aligned}$$

Here we impose the weight function to have the (deterministic) structure  $w_n = \frac{\pi(\mathbf{x}_n)}{\varphi_{\mathcal{P}_n}(\mathbf{x}_n)}$ , where  $\pi(\mathbf{x}_n)$  is the target density and  $\varphi_n(\mathbf{x}_n)$  is a generic function parametrized by a set of parameters  $\mathcal{P}_n$ , and both terms are evaluated at  $\mathbf{x}_n$ .<sup>8</sup> The expectation of the generic estimator can be then computed as

$$E[\hat{I}] = \frac{1}{ZN} \sum_{n=1}^N \sum_{j_{1:N}} \int \frac{\pi(\mathbf{x}_n) g(\mathbf{x}_n)}{\varphi_{\mathcal{P}_n}(\mathbf{x}_n)} P(j_{1:N}) p(\mathbf{x}_n | j_n) d\mathbf{x}_n. \quad (26)$$

**Remark 2. (Weighting):** In the proposed framework, we consider valid any weighting scheme (i.e., any function  $\varphi_{\mathcal{P}_n}$  at the denominator of the weight) that yields  $E[\hat{I}] \equiv I$  in Eq. (26).

In the following, we show that various distributions related to the generation of the samples (discussed in Section III) can be used as the denominator of the weight, yielding valid estimators.

### B. Weighting functions

Here we present several possible functions  $\varphi_{\mathcal{P}_n}$ , that yield an unbiased estimator of  $I$  according to Eq. (26). The different choices for  $\varphi_{\mathcal{P}_n}$  come naturally from the sampling densities discussed in the previous section. More precisely, they correspond to the appropriate evaluation at  $\mathbf{x}_n$  of the five different functions in Table I related to the distributions of the generated samples. From now on,  $p(\cdot)$  and  $f(\cdot)$ , which correspond to the pdfs of  $\mathbf{X}_n$  and  $\mathbf{X}$  respectively, are used as functions and the argument represents a functional evaluation.

$$\mathcal{W}_1: \varphi_{\mathcal{P}_n}(\mathbf{x}_n) = \varphi_{j_{1:n-1}}(\mathbf{x}_n) = p(\mathbf{x}_n | j_{1:n-1})$$

Since the sampling process is sequential, this option is of particular interest. It interprets the proposal pdf as the conditional density of  $\mathbf{x}_n$  given all the previous proposal indexes of the sampling process.

$$\mathcal{W}_2: \varphi_{\mathcal{P}_n}(\mathbf{x}_n) = \varphi_{j_n}(\mathbf{x}_n) = p(\mathbf{x}_n | j_n) = q_{j_n}(\mathbf{x}_n)$$

It interprets that if the index  $j_n$  is known,  $\varphi$  is the proposal  $q_{j_n}$ .

$$\mathcal{W}_3: \varphi_{\mathcal{P}_n}(\mathbf{x}_n) = p(\mathbf{x}_n)$$

It interprets that  $\mathbf{x}_n$  is a realization of the marginal  $p(\mathbf{x}_n)$ . This is probably the most “natural” option (as it does not assume any further knowledge in the generation of  $\mathbf{x}_n$ ) and is a usual choice for the calculation of the weights in some of the existing MIS schemes (see Section V).

$$\mathcal{W}_4: \varphi_{\mathcal{P}_n}(\mathbf{x}_n) = \varphi_{j_{1:N}}(\mathbf{x}_n) = f(\mathbf{x}_n | j_{1:N}) = \frac{1}{N} \sum_{k=1}^N q_{j_k}(\mathbf{x}_n)$$

This interpretation makes use of the distribution of the r.v.  $\mathbf{X}$  conditioned on the whole set of indexes (defined in Section III-D).

$$\mathcal{W}_5: \varphi_{\mathcal{P}_n}(\mathbf{x}_n) = \varphi(\mathbf{x}_n) = f(\mathbf{x}_n) = \frac{1}{N} \sum_{k=1}^N q_k(\mathbf{x}_n)$$

<sup>8</sup>Note that, in an even more generalized framework, the  $n$ -th weight  $w_n$  could hypothetically depend on more than one sample of the set  $\mathbf{x}_{1:N}$  if one could properly design the function  $\varphi_n$  that yields valid estimators.

TABLE II: Summary of the different generic functions  $\varphi_{\mathcal{P}_n}$ . The distributions depend on the specific sampling scheme used for drawing the samples as shown in Table III.

| $\varphi_{\mathcal{P}_n}$   | $\mathcal{W}_1$<br>$p(\mathbf{x}_n   j_{1:n-1})$        | $\mathcal{W}_2$<br>$p(\mathbf{x}_n   j_n)$        | $\mathcal{W}_3$<br>$p(\mathbf{x}_n)$        | $\mathcal{W}_4$<br>$f(\mathbf{x}   j_{1:N})$          | $\mathcal{W}_5$<br>$f(\mathbf{x})$          |
|---|---|---|---|---|---|
| $w_n = \frac{\pi(\mathbf{x}_n)}{\varphi_{\mathcal{P}_n}(\mathbf{x}_n)}$ | $\frac{\pi(\mathbf{x}_n)}{p(\mathbf{x}_n   j_{1:n-1})}$ | $\frac{\pi(\mathbf{x}_n)}{p(\mathbf{x}_n   j_n)}$ | $\frac{\pi(\mathbf{x}_n)}{p(\mathbf{x}_n)}$ | $\frac{\pi(\mathbf{x}_n)}{f(\mathbf{x}_n   j_{1:N})}$ | $\frac{\pi(\mathbf{x}_n)}{f(\mathbf{x}_n)}$ |

TABLE III: Specific function,  $\varphi_{\mathcal{P}_n}$ , at the denominator of weight,  $w_n = \frac{\pi(\mathbf{x}_n)}{\varphi_{\mathcal{P}_n}(\mathbf{x}_n)}$ , resulting from the combination of the different sampling schemes (Section III-D) and weighting functions (Section IV-B).

| $\varphi_{\mathcal{P}_n}$                    | $\mathcal{W}_1$<br>$p(\mathbf{x}_n   j_{1:n-1})$                                      | $\mathcal{W}_2$<br>$p(\mathbf{x}_n   j_n)$ | $\mathcal{W}_3$<br>$p(\mathbf{x}_n)$ | $\mathcal{W}_4$<br>$f(\mathbf{x}   j_{1:N})$          | $\mathcal{W}_5$<br>$f(\mathbf{x})$ |
|--|---|--|--------------------------------------|---|------------------------------------|
| $\mathcal{S}_1$ : <b>with replacement</b>    | $\psi(\mathbf{x}_n)$ [R3]   | $q_{j_n}(\mathbf{x}_n)$ [R1]               | $\psi(\mathbf{x}_n)$ [R3]            | $\frac{1}{N} \sum_{k=1}^N q_{j_k}(\mathbf{x}_n)$ [R2] | $\psi(\mathbf{x}_n)$ [R3]          |
| $\mathcal{S}_2$ : <b>w/o (random)</b>        | $\frac{1}{ \mathcal{I}_n } \sum_{\forall k \in \mathcal{I}_n} q_k(\mathbf{x}_n)$ [N2] | $q_{j_n}(\mathbf{x}_n)$ [N1]               | $\psi(\mathbf{x}_n)$ [N3]            | $\psi(\mathbf{x}_n)$ [N3]                             | $\psi(\mathbf{x}_n)$ [N3]          |
| $\mathcal{S}_3$ : <b>w/o (deterministic)</b> | $q_n(\mathbf{x}_n)$ [N1]  | $q_n(\mathbf{x}_n)$ [N1]                   | $q_n(\mathbf{x}_n)$ [N1]             | $\psi(\mathbf{x}_n)$ [N3]                             | $\psi(\mathbf{x}_n)$ [N3]          |

This option considers that all the  $\mathbf{x}_n$  are realizations of the r.v.  $\mathbf{X}$  defined in Section III-D (see Appendix A for a thorough discussion of this interpretation).

Although some of the selected functions  $\varphi_{\mathcal{P}_n}$  may seem more natural than others, all of them yield valid estimators. The proofs can be found in Appendix A. Table II summarizes the discussed functions  $\varphi_{\mathcal{P}_n}$  that can be used to evaluate the denominator for the weight calculation,  $w_n = \frac{\pi(\mathbf{x}_n)}{\varphi_{\mathcal{P}_n}(\mathbf{x}_n)}$ . Other proper weighting functions are described in Section VII-B.

## V. MIS SCHEMES

In this section, we describe the different possible combinations of the sampling strategies considered in Section III) and the weighting functions devised in Section IV. We note that, even though we have discussed three sampling procedures and five alternatives for weight calculation, once combined the fifteen possibilities only lead to six unique MIS methods. Three of the methods are associated to the sampling scheme with replacement ( $\mathcal{S}_1$ ), while the other three methods correspond to the sampling schemes without replacement ( $\mathcal{S}_2$  and  $\mathcal{S}_3$ ). Note that for each specific sampling (i.e., with or without replacement), different weighting options can yield the same function in the denominator (e.g. for the deterministic sampling without replacement,  $\mathcal{S}_3$ , the denominators for weighting options 1, 2 and 3 are identical, always yielding  $w_n = \frac{\pi(\mathbf{x}_n)}{q_n(\mathbf{x}_n)}$ ). Table III summarizes the possible combinations of sampling/weighting and indicates the resulting MIS method within brackets. The six MIS methods are labeled either by an R indicating that the method uses sampling with *replacement* or with an N to denote that the method corresponds to a sampling scheme with *no* replacement. We remark that these schemes are examples of proper MIS techniques fulfilling Remarks 1 and 2.

### A. MIS schemes with replacement

In all R schemes, the  $n$ -th sample is drawn with replacement (i.e.,  $\mathcal{S}_1$ ) from the whole mixture  $\psi$ :

[R1]: *Sampling with replacement,  $\mathcal{S}_1$ , and weight denominator  $\mathcal{W}_2$ :*

For the weight calculation of the  $n$ -th sample, only the mixand selected for generating the sample is evaluated in the denominator.

[R2]: *Sampling with replacement,  $\mathcal{S}_1$ , and weight denominator  $\mathcal{W}_4$ :*

With the  $N$  selected indexes  $j_n$ , for  $n = 1, \dots, N$ , one forms a mixture composed by all the corresponding proposal

pdfs. The weight calculation of the  $n$ -th sample considers this *a posteriori* mixture evaluated at the  $n$ -th sample in the denominator, i.e., some proposals might be used more than once while other proposals might not be used.

[R3]: *Sampling with replacement,  $\mathcal{S}_1$ , and weight denominator  $\mathcal{W}_1$ ,  $\mathcal{W}_3$ , or  $\mathcal{W}_5$ :*

For the weight calculation of the  $n$ -th sample, the denominator applies the value of the  $n$ -th sample to the whole mixture  $\psi$  composed of the set of initial proposal pdfs (i.e., the function in the denominator of the weight does not depend on the sampling process). This is the approach followed by the so called mixture PMC method [9].

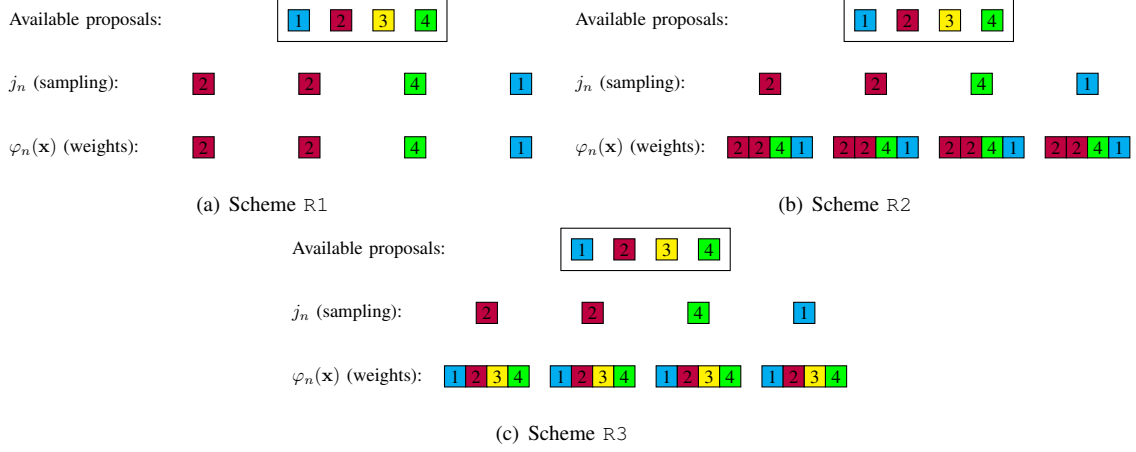


Fig. 3: Example of a realization of the indexes selection ( $N = 4$ ) with the procedure  $\mathcal{S}_1$  (with replacement), and the corresponding possible denominators for weighting: (a) [R1]:  $\varphi_n(\mathbf{x}_n) = q_{j_n}(\mathbf{x}_n)$ ; (b) [R2]:  $\varphi_n(\mathbf{x}_n) = \frac{1}{N} \sum_{k=1}^N q_{j_k}(\mathbf{x}_n)$ ; (c) [R3]:  $\varphi_n(\mathbf{x}_n) = \psi(\mathbf{x}_n) = \frac{1}{N} \sum_{k=1}^N q_k(\mathbf{x}_n)$ .

Figure 3 displays these three MIS schemes related to the sampling with replacement,  $\mathcal{S}_1$ . It shows a possible realization of the MIS schemes for  $M = N = 4$  samples and pdfs. For each scheme, this figure depicts for each  $n$ -th sample, the index  $j_n$  of the proposal pdf that was actually selected to draw the sample, and the function  $\varphi_n$  that was used for the denominator of the weight calculation. Note that all these MIS schemes allow for repetition in the selected proposal pdf (since the selection of proposal pdfs is done with replacement). In Fig. 3(a), the denominator of the weight calculation only considers the evaluation of the actual proposal pdf from which the sample was generated (scheme R1). In Fig. 3(b) the denominator of the weight accounts for the mixture of proposal pdfs selected, but evaluated at the  $n$ -th sample value (scheme R2). Finally, in Fig. 3(c), the weight evaluation uses the complete mixture  $\psi$  of initial proposal pdfs evaluated at the sample value (scheme R3).

### B. MIS schemes without replacement

In all N schemes, exactly one sample is generated from each proposal pdf. This corresponds to having a sampling strategy without replacement.

[N1]: *Sampling without replacement (random or deterministic),  $\mathcal{S}_2$  or  $\mathcal{S}_3$ , and weight denominator  $\mathcal{W}_2$  (for  $\mathcal{S}_2$ ) or  $\mathcal{W}_1$ ,  $\mathcal{W}_2$ , or  $\mathcal{W}_3$  (for  $\mathcal{S}_3$ ):*

For calculating the denominator of the  $n$ -th weight, the specific mixand used for the generation of the sample is used. This is the approach frequently used in particle filtering [13] and in the standard PMC method [8].

[N2]: *Sampling without replacement (random),  $\mathcal{S}_2$ , and weight denominator  $\mathcal{W}_1$ :*

This MIS implementation draws one sample from each mixand, but the order matters (it must be random) since the

calculation of the  $n$ -th weight uses for the evaluation of the denominator the mixture pdf formed by the proposal pdfs that were still available at the generation of the  $n$ -th sample.

[N3]: *Sampling without replacement (random or deterministic),  $\mathcal{S}_2$  or  $\mathcal{S}_3$ , and weight denominator  $\mathcal{W}_3$ ,  $\mathcal{W}_4$ , or  $\mathcal{W}_5$  (for  $\mathcal{S}_2$ ), or  $\mathcal{W}_4$  or  $\mathcal{W}_5$  (for  $\mathcal{S}_3$ ):*

In the calculation of the  $n$ -th weight, one uses for the denominator the whole mixture. This is the approach, for instance, of [10, 11]. As showed in Section VI, this scheme has several benefits over the others.

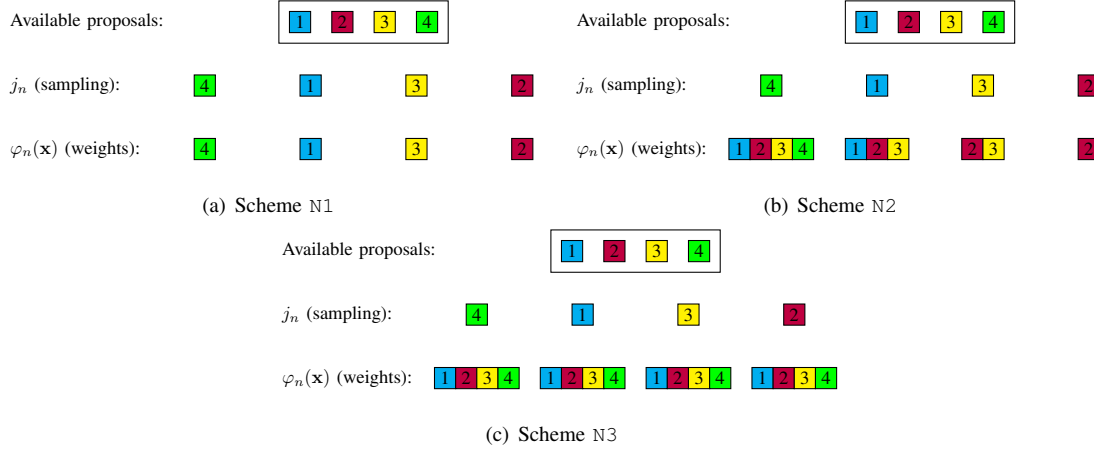


Fig. 4: Example of a realization of the indexes selection ( $N = 4$ ) with the procedure  $\mathcal{S}_2$  (without replacement), and the corresponding possible denominators for weighting: (a) [N1]:  $\varphi_n(\mathbf{x}_n) = q_{j_n}(\mathbf{x}_n)$ ; (b) [N2]:  $\varphi_n(\mathbf{x}_n) = \frac{1}{|\mathcal{I}_n|} \sum_{\forall k \in \mathcal{I}_n} q_k(\mathbf{x}_n)$ ; (c) [N3]:  $\varphi_n(\mathbf{x}_n) = \frac{1}{N} \sum_{k=1}^N q_k(\mathbf{x}_n)$ .

Figures 4 and 6 display the three MIS schemes related to the sampling without replacement,  $\mathcal{S}_2$  and  $\mathcal{S}_3$ , respectively. Note that the sampling always yields one sample from each of the proposal pdfs of the available set. In Fig. 4 (a), the denominator of the weight calculation only considers the evaluation of the actual proposal pdf from which the sample was generated (scheme N1). In Fig. 4(b), the denominator of the weight accounts for the mixture of proposal pdfs available when the  $n$ -th sample was generated and evaluated at the  $n$ -th sample value (scheme N2). Finally, in Fig. 4(c) the weight evaluation uses the complete mixture of initial proposal pdfs evaluated at the sample value (scheme N3). We can see that schemes in Figs. 4 (a) and 6 (a) are equivalent. The MIS schemes of Fig. 4 (c) and Fig. 6 (b) are also equivalent. In both cases, this equivalency is due to the fact that the order does not play any role in the construction of the estimators or the sample approximation of the target density.

Figure 5 illustrates the possible proper combinations of sampling procedures and weighting functions. Table IV summarizes the six resulting MIS schemes and their references in literature, indicating which sampling procedure and weighting function must be applied to obtain the  $n$ -th weighted sample  $\mathbf{x}_n$ . We consider N1 and N3 associated to  $\mathcal{S}_3$  (they can also be obtained with  $\mathcal{S}_2$ ) since it is a simpler sampler than  $\mathcal{S}_2$ .

Within the proposed framework, we have considered three sampling procedures and five general weighting functions. All the different algorithms in the literature (that we are aware of) correspond to one of the MIS schemes described above. Section VII-C provides more details about the MIS schemes used by the different algorithms available in literature. Several new valid schemes have also appeared naturally. Namely, schemes R1, R2, and N2 are novel, and their advantages and drawbacks are analyzed in the following sections. Furthermore, following the sampling and weighting remarks provided above, new proper MIS schemes can easily be proposed within this framework.

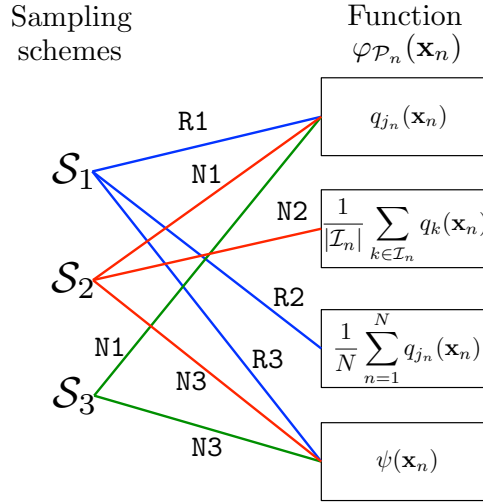


Fig. 5: Proper combinations of sampling procedures and weighting functions applied to the  $n$ -th sample  $\mathbf{x}_n$ .

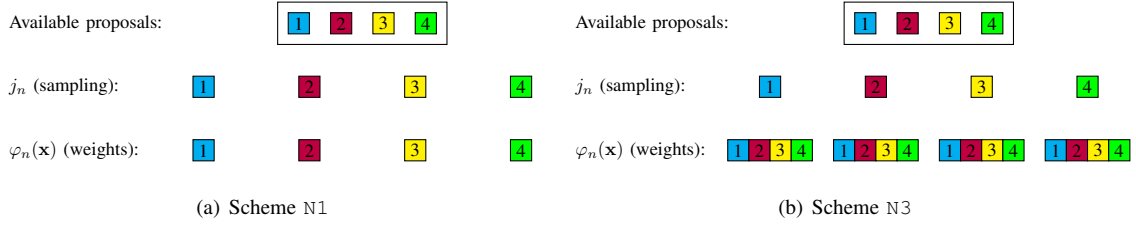


Fig. 6: Indexes selection ( $N = 4$ ) with the procedure  $\mathcal{S}_3$  (without replacement), and the corresponding possible denominators for weighting: (a)  $\varphi_n(\mathbf{x}_n) = q_{j_n}(\mathbf{x}_n)$ ; (b)  $\varphi_n(\mathbf{x}_n) = \psi(\mathbf{x}_n) = \frac{1}{N} \sum_{k=1}^N q_k(\mathbf{x}_n)$ .

TABLE IV: Summary of the sampling procedure and the weighting function of each MIS scheme.

| MIS scheme | Sampling        | $w(\mathbf{x}_n)$  | Used in  |
|------------|-----------------|--|----------|
| R1         | $\mathcal{S}_1$ | $\frac{\pi(\mathbf{x}_n)}{q_{j_n}(\mathbf{x}_n)}$  | Novel    |
| R2         | $\mathcal{S}_1$ | $\frac{\pi(\mathbf{x}_n)}{\frac{1}{N} \sum_{k=1}^N q_{j_k}(\mathbf{x}_n)}$                         | Novel    |
| R3         | $\mathcal{S}_1$ | $\frac{\pi(\mathbf{x}_n)}{\psi(\mathbf{x}_n)}$   | [9]      |
| N1         | $\mathcal{S}_3$ | $\frac{\pi(\mathbf{x}_n)}{q_n(\mathbf{x}_n)}$  | [8]      |
| N2         | $\mathcal{S}_2$ | $\frac{\pi(\mathbf{x}_n)}{\frac{1}{ \mathcal{I}_n } \sum_{k \in \mathcal{I}_n} q_k(\mathbf{x}_n)}$ | Novel    |
| N3         | $\mathcal{S}_3$ | $\frac{\pi(\mathbf{x}_n)}{\psi(\mathbf{x}_n)}$   | [10, 11] |

## VI. QUALITY MEASURES OF THE SCHEMES

In this section we analyze the MIS schemes presented in the previous section in terms of objective quality measures. In particular, we provide an exhaustive variance analysis. Moreover, a discussion about the effective sample size and the application of resampling procedures is provided.

### A. Variance Analysis

A central objective in importance sampling entails the computation of a particular moment of r.v. with pdf  $\tilde{\pi}(\mathbf{x}) = \frac{\pi(\mathbf{x})}{Z}$ . For sake of completeness of this section, let us revisit the general forms of the estimators. We recall that the goal is approximating

$$I = \frac{1}{Z} \int g(\mathbf{x}) \pi(\mathbf{x}) d\mathbf{x}, \quad (27)$$

where  $g$  can be any square integrable function of  $\mathbf{x}$  [2].

In standard importance sampling, the moment in Eq. (27) can be estimated by drawing  $N$  independent samples  $\mathbf{x}_n$  from a single proposal density  $q(\mathbf{x})$  and building the estimator

$$\tilde{I} = \frac{1}{N\hat{Z}} \sum_{n=1}^N w_n g(\mathbf{x}_n), \quad (28)$$

where  $w_n = \frac{\pi(\mathbf{x}_n)}{q(\mathbf{x}_n)}$  for  $n = 1, \dots, N$ , and  $\hat{Z} = \frac{1}{N} \sum_{j=1}^N w_j$ . Under mild assumptions about the tails of the distributions,  $\hat{I}$  provides a consistent estimator of  $I$  [1]. If the normalizing constant  $Z$  of the target  $\pi(\mathbf{x})$  is known, the estimator

$$\hat{I} = \frac{1}{NZ} \sum_{n=1}^N w_n g(\mathbf{x}_n), \quad (29)$$

is also unbiased [1, 2]. Furthermore, it is well known that the variance of both estimators is directly related to the discrepancy between  $\tilde{\pi}(\mathbf{x})|g(\mathbf{x})|$  and  $q(\mathbf{x})$  (for a specific choice of  $g$ ) [1, 14]. For a general  $g$ , a common strategy is decreasing the mismatch between the proposal  $q(\mathbf{x})$  and the target  $\tilde{\pi}(\mathbf{x})$ .

In MIS, a set of  $N$  proposal pdfs  $\{q_n(\mathbf{x})\}_{n=1}^N$  is used to draw the  $N$  samples. While the MIS estimators preserve the same structure as in Eqs. (28) and (29), the way the samples are drawn (see the sampling procedures in Section III) and the function used for the weight calculation (see Section IV) can make a substantial difference in the performance. In fact, although the six different MIS schemes that appear in Section V yield an unbiased estimator (see Appendix A), the performance of that estimator can be dramatically different. In the following, we focus on the variance of the unbiased estimator  $\hat{I}$  of Eq. (29) in all the studied schemes. The details of the derivations are in Appendix B. In particular, the estimators of the three methods with replacement present the following variances

$$\text{Var}(\hat{I}_{R1}) = \frac{1}{Z^2 N^2} \sum_{k=1}^N \int \frac{\pi^2(\mathbf{x}) g^2(\mathbf{x})}{q_k(\mathbf{x})} d\mathbf{x} - \frac{I^2}{N}, \quad (30)$$

$$\begin{aligned} \text{Var}(\hat{I}_{R2}) &= \frac{1}{Z^2 N} \frac{1}{N^N} \sum_{j_{1:N}} \int \frac{\pi^2(\mathbf{x}) g^2(\mathbf{x})}{f(\mathbf{x}|j_{1:N})} d\mathbf{x} \\ &\quad - \frac{1}{Z^2 N^2} \frac{1}{N^N} \sum_{j_{1:N}} \sum_{n=1}^N \left( \int \frac{\pi(\mathbf{x}_n) g(\mathbf{x}_n)}{f(\mathbf{x}_n|j_{1:N})} q_{j_n}(\mathbf{x}_n) d\mathbf{x}_n \right)^2, \end{aligned} \quad (31)$$

and

$$\text{Var}(\hat{I}_{R3}) = \frac{1}{Z^2 N} \int \frac{\pi^2(\mathbf{x}) g^2(\mathbf{x})}{\psi(\mathbf{x})} d\mathbf{x} - \frac{I^2}{N}. \quad (32)$$

On the other hand, the variances associated to the estimators of the three methods without replacement are

$$\text{Var}(\hat{I}_{N1}) = \frac{1}{Z^2 N^2} \sum_{n=1}^N \int \frac{\pi^2(\mathbf{x}_n) g^2(\mathbf{x}_n)}{q_n(\mathbf{x}_n)} d\mathbf{x}_n - \frac{I^2}{N}, \quad (33)$$

$$\begin{aligned} \text{Var}(\hat{I}_{N2}) &= \left[ \frac{1}{Z^2 N^2} \sum_{n=1}^N \sum_{j_{1:n-1}} \int \frac{\pi^2(\mathbf{x}_n) g^2(\mathbf{x}_n)}{p(\mathbf{x}_n|j_{1:n-1})} P(j_{1:n-1}) d\mathbf{x}_n \right] \\ &\quad - \left[ \frac{1}{Z^2 N^2} \sum_{n=1}^N \sum_{j_{1:n}} \left( \int \frac{\pi(\mathbf{x}_n) g(\mathbf{x}_n)}{p(\mathbf{x}_n|j_{1:n-1})} q_{j_n} d\mathbf{x}_n \right)^2 \right] P(j_{1:n}), \end{aligned} \quad (34)$$

and

$$\begin{aligned} \text{Var}(\hat{I}_{N3}) &= \frac{1}{Z^2 N} \int \frac{\pi^2(\mathbf{x}) g^2(\mathbf{x})}{\psi(\mathbf{x})} d\mathbf{x} \\ &\quad - \frac{1}{Z^2 N^2} \sum_{n=1}^N \left( \int \frac{\pi(\mathbf{x}) g(\mathbf{x})}{\psi(\mathbf{x})} q_n(\mathbf{x}) d\mathbf{x} \right)^2. \end{aligned} \quad (35)$$

One of the goals of this paper is to provide the practitioner with solid theoretical results about the superiority of some specific MIS schemes. In the following, we state two theorems that relate the variance of the estimator with these six methods, establishing a hierarchy among them. Note that obtaining an IS estimator with finite variance essentially amounts to having a proposal with heavier tails than the target. See [1, 15] for sufficient conditions that guarantee this finite variance.

**Theorem 1.** *For any target distribution  $\pi(\mathbf{x})$ , any square integrable function  $g$ , and any set of proposal densities  $\{q_n(\mathbf{x})\}_{n=1}^N$  such that the variance of the corresponding MIS estimators is finite,*

$$\text{Var}(\hat{I}_{R1}) = \text{Var}(\hat{I}_{N1}) \geq \text{Var}(\hat{I}_{R3}) \geq \text{Var}(\hat{I}_{N3})$$

**Proof:** See Appendix B. □

**Theorem 2.** *For any target distribution  $\pi(\mathbf{x})$ , any square integrable function  $g$ , and any set of proposal densities  $\{q_n(\mathbf{x})\}_{n=1}^2$  such that the variance of the corresponding MIS estimators is finite,*

$$\text{Var}(\hat{I}_{R1}) = \text{Var}(\hat{I}_{N1}) \geq \text{Var}(\hat{I}_{R2}) = \text{Var}(\hat{I}_{N2}) \geq \text{Var}(\hat{I}_{N3}) \quad (36)$$

**Proof:** See Appendix C. □

First, let us note that the scheme N3 outperforms (in terms of the variance) any other MIS scheme in the literature that we are aware of. Moreover, for  $N = 2$ , it also outperforms the other novel schemes R2 and N2. While the MIS schemes R2 and N2 do not appear in Theorem 1, we hypothesize that the conclusions of Theorem 2 might be extended to  $N > 2$ . The intuitive reason is that, regardless of  $N$ , both methods partially reduce the variance of the estimators by placing more than one proposal at the denominator of some or all the weights.

The variance analysis of  $\tilde{I}$  in Eq. (28) implies a ratio of dependent r.v.'s, and therefore, it cannot be performed without resorting to an approximation, e.g., by means of a Taylor expansion as it is performed in [16, 17, 3]. In this case, the bias of  $\tilde{I}$  is usually considered negligible compared to the variance for large  $N$ . With this approximation, the variance depends on the variances of the numerator (which is a scaled version of  $\hat{I}$ ), the variance of  $\hat{Z}$ , and the covariance of both. Therefore, although we prove several relations in terms of the variance for  $\hat{I}$  and  $\hat{Z}$ , the same conclusions for the normalized estimator  $\hat{I}_{\text{MIS}}$  cannot strictly be proved in the general case. However, it is reasonable to assume that methods that reduce the variance of  $\hat{I}$  and  $\hat{Z}$ , in general will also reduce the variance of  $\tilde{I}$ . In Section VIII, this hypothesis is reinforced by means of numerical simulations.

### B. Effective Sample Size

Let us consider again the problem of estimating the integral of Eq. (27). If one could draw  $N$  independent samples directly from the normalized target distribution,  $\mathbf{X}_n \sim \tilde{\pi}(\mathbf{x}_n)$  for  $n = 1, \dots, N$ , the estimator of  $I$  would be given by

$$\bar{I} = \frac{1}{N} \sum_{n=1}^N g(\mathbf{x}_n). \quad (37)$$



The efficiency of an importance sampling method is often measured with the so-called effective sample size (ESS). In its original formulation, the ESS measures the increase of variance w.r.t. to the estimator of Eq. (37) [16, 17]. In standard importance sampling (i.e., with a single proposal), the ESS of the IS estimator of Eq. (28) is approximated as [16, 17]

$$ESS_{\text{IS}} = \frac{N \text{Var}_q[\tilde{I}]}{\text{Var}_{\tilde{\pi}}[\tilde{I}]} \approx \frac{N}{1 + \text{Var}_q[w]}. \quad (38)$$

An appropriate extension of Eq. (38) to MIS is not straightforward, since the samples are drawn from different proposal pdfs. Noting that in standard IS  $\text{Var}_q[\hat{Z}] = \frac{\text{Var}_q[w]}{N}$ , we propose the following natural extension of the ESS to the MIS approach:

$$ESS_{\text{MIS}} \approx \frac{N}{1 + N \text{Var}[\hat{Z}]}, \quad (39)$$

where the variance of  $\hat{Z}$  is now calculated over the joint distribution  $p(\mathbf{x}_{1:N}, j_{1:N})$ . In this way, methods with a lower value of  $\text{Var}[\hat{Z}]$  present a higher value of ESS. Therefore, the analysis of the variance of the generic estimator  $\hat{I}$  (Section VI-A) can be straightforwardly applied here, yielding the same hierarchy among the different MIS schemes.

Since the variances involved in the ESS calculation cannot usually be obtained in a closed form, practitioners in Sequential Monte Carlo methods often use an approximation of the variance. In particular, as a rule of thumb, the following expression is commonly used

$$\widehat{ESS} = \frac{1}{\sum_{n=1}^N \bar{w}_n^2}, \quad (40)$$

where  $\bar{w}_n = \frac{w_n}{\sum_{k=1}^N w_k}$  is the  $n$ -th normalized weight [18, 19]. Note that  $\widehat{ESS} = N$  (maximum value) when all weights are equal, whereas  $\widehat{ESS} = 1$  (minimum value) when all weights are zero except one.

Note that MIS schemes with lower  $E[\sum_{n=1}^N \bar{w}_n^2]$  will yield a bigger  $\widehat{ESS}$  in average. Therefore, we have

$$\begin{aligned} \sum_{n=1}^N E[\bar{w}_n^2] &= \sum_{n=1}^N E \left[ \frac{w_n^2}{\left( \sum_{k=1}^N w_k \right)^2} \right] \\ &= \frac{1}{N^2} \sum_{n=1}^N E \left[ \frac{w_n^2}{\hat{Z}^2} \right]. \end{aligned} \quad (41)$$

It is not straightforward to analyze this ratio of dependent r.v. However, since  $\hat{Z} \approx Z$  for large  $N$ , we can consider that MIS schemes with smaller  $\sum_{n=1}^N E[w_n^2]$  have bigger  $\widehat{ESS}$  in average. In particular, regarding the proof of Theorem 1 at the variance analysis, and under the assumption of  $\hat{Z} = Z$ , we can see that the MIS schemes R3 and N3 outperform R1 and N1 (compare the first terms of Eqs. (30)-(35) and see the Appendix A for more details). Nevertheless, this interpretation must be taken carefully, since it comes from several approximations and assumptions. Following these lines, in [3, Chapter 9], is stated that  $\widehat{ESS}$  “is a convenient way to interpret the amount of inequality in the weights but it does not translate cleanly into a comparison of variances”.

In addition, we recall that the original effective sample size is defined as a ratio (scaled by  $N$ ) that measures the loss in terms of variance of the Importance Sampling estimator w.r.t. to the ideal estimator  $\bar{I}$ . In some specific cases, the estimator of the MIS scheme may have less variance than in direct target sampling (see the second example of Section VIII). Therefore,  $\widehat{ESS}$  of Eq. (40), which is derived through several approximations and assumptions, might not always be capturing the essence of the effective sample size.

### C. Resampling in Adaptive Importance Sampling

Many adaptive importance sampling (AIS) algorithms, such as the standard PMC [8] or sequential Monte Carlo methods [13, 20], employ resampling steps for updating the parameters of the proposal functions. In the standard multinomial resampling,

the indexes of the next generation are drawn according to a multinomial pmf parameterized by the normalized IS weights  $\bar{w}_n$ , with  $n = 1, \dots, N$ . One of the main issues in the application of resampling steps is the loss of diversity, also known as path degeneracy, in the samples of the next generation. This loss is minimized when the pmf described by the normalized weights  $\bar{w}_n = \frac{w_n}{N\hat{Z}}$ , with  $n = 1, \dots, N$ , coincides with the discrete uniform pmf  $\mathcal{U}\{1, 2, \dots, N\}$ , i.e.,  $\bar{w}_n = \frac{1}{N}$  for all  $n$ . Thus, let us consider the Euclidean distance between these two pmf's, i.e.,

$$\begin{aligned} L_2 &= \sum_{n=1}^N \left( \frac{w_n}{\sum_{k=1}^N w_k} - \frac{1}{N} \right)^2 \\ &= \frac{1}{N^2} \frac{1}{\hat{Z}^2} \sum_{n=1}^N \left( w_n - \hat{Z} \right)^2, \end{aligned} \quad (42)$$

where we have used  $\hat{Z} = \frac{1}{N} \sum_{k=1}^N w_k$ . With simple rearrangements, we have

$$\begin{aligned} L_2 &= \frac{1}{N^2} \frac{1}{\hat{Z}^2} \left[ \sum_{n=1}^N w_n^2 + N\hat{Z}^2 - 2\hat{Z} \sum_{n=1}^N w_n \right] \\ &= \frac{1}{N^2} \frac{1}{\hat{Z}^2} \left[ \sum_{n=1}^N w_n^2 - N\hat{Z}^2 \right] \\ &= \sum_{n=1}^N \bar{w}_n^2 - \frac{1}{N} = \frac{1}{\widehat{ESS}} - \frac{1}{N}, \end{aligned} \quad (43)$$

where  $\widehat{ESS} = \frac{1}{\sum_{n=1}^N \bar{w}_n^2}$  (see Eq. (40)). Therefore, maximizing the  $\widehat{ESS}$  is equivalent to minimizing the Euclidean distance between the pmf of the resampled index (parametrized by the unnormalized weights) and the discrete uniform pmf. As a consequence, a MIS scheme that provides a higher  $\widehat{ESS}$ , also provides a pmf closer to the uniform distribution. Therefore, from the point of view of reducing the discrepancy of the weights and retaining more diversity after a resampling step, we can again state that MIS schemes R3 and N3 outperform R1 and N1.

## VII. APPLYING THE MIS SCHEMES

### A. Computational complexity

In previous section, we have compared the MIS schemes in terms of performance, whereas here we discuss their computational complexity. Table V compares the total number of target and proposal evaluations in each MIS scheme. First, note that the estimators of any MIS scheme within the proposed general framework use  $N$  weighted samples where the general weight is given by Eq. (20).<sup>9</sup> Therefore, all of them perform  $N$  target evaluations in total. However, depending on the function  $\varphi_{\mathcal{P}_n}$  used by each specific scheme at the weight denominator, a different number of proposal evaluations is performed. We see that R3, and N3 always yield the largest number of proposal evaluations. In R2, the number of proposal evaluations is variable: although each weight evaluates  $N$  proposals, some proposals may be repeated, whereas others may not be used.

In many relevant scenarios, the cost of evaluating the proposal densities is negligible compared to the cost of evaluating the target function. In this scenario, the MIS scheme N3 should always be chosen, since it yields a lower variance with a negligible increase in computational cost. For instance, this is the case in the *Big Data* Bayesian framework, where the target function is a posterior distribution with many data in the likelihood function. However, in some other scenarios, e.g. when the number of proposals  $N$  is too large and/or the target evaluations are not very expensive, limiting the number of proposal evaluations can result in a better cost-performance trade off.

<sup>9</sup>We recall that, in general, one can draw  $M = kN$  samples, with  $k \geq 1$  and  $k \in \mathbb{N}$ .

| MIS Scheme           | R1  | N1  | R2         | N2         | R3    | N3    |
|----------------------|-----|-----|------------|------------|-------|-------|
| Target Evaluations   | $N$ | $N$ | $N$        | $N$        | $N$   | $N$   |
| Proposal Evaluations | $N$ | $N$ | $\leq N^2$ | $N(N+1)/2$ | $N^2$ | $N^2$ |

TABLE V: Number of target and proposal evaluations for the different MIS schemes. Note that the number of proposal evaluations for R2 is a random variable with a range from  $N$  to  $N^2$ .

### B. A priori partition approach

The extra computational cost of some MIS schemes occurs because each sample must be evaluated in more than one proposal  $q_n$ , or even in all of the available proposals (e.g. the MIS scheme N3). In order to propose a framework that limits the number of proposal evaluations, let us first define a partition of the set of the indexes of all proposals,  $\{1, \dots, N\}$ , into  $P$  disjoint subsets of  $L$  elements (indexes),  $\mathcal{J}_p$  with  $p = 1, \dots, P$ , s.t.

$$\mathcal{J}_1 \cup \mathcal{J}_2 \cup \dots \cup \mathcal{J}_P = \{1, \dots, N\}, \quad (44)$$

where  $\mathcal{J}_k \cap \mathcal{J}_q = \emptyset$  for all  $k, q = 1, \dots, P$  and  $k \neq q$ .<sup>10</sup> Therefore, each subset,  $\mathcal{J}_p = \{j_{p,1}, j_{p,2}, \dots, j_{p,L}\}$ , contains  $L$  indexes,  $j_{p,\ell} \in \{1, \dots, N\}$  for  $\ell = 1, \dots, L$  and  $p = 1, \dots, P$ .

After this *a priori* partition, one could apply any MIS scheme in each (partial) subset of proposals, and then perform a suitable convex combination of the partial estimators. This general strategy is inspired by a specific scheme, partial deterministic mixture MIS (p-DM-MIS), which was recently proposed in [7]. That work applies the idea of the partitions just for the MIS scheme N3, denoted there as full deterministic mixture MIS (f-DM-MIS). The sampling procedure is then  $\mathcal{S}_3$ , i.e., exactly one sample is drawn from each proposal. The weight of each sample in p-DM-MIS, instead of evaluating the whole set of proposals (as in N3), evaluates only the proposals within the subset that the generating proposal belongs to. Mathematically, the weights of the samples corresponding to the  $p$ -th mixture are computed as

$$w_n = \frac{\pi(\mathbf{x}_n)}{\psi_p(\mathbf{x}_n)} = \frac{\pi(\mathbf{x}_n)}{\frac{1}{L} \sum_{j \in \mathcal{J}_p} q_j(\mathbf{x}_n)}, \quad n \in \mathcal{J}_p. \quad (45)$$

Note that the number of proposal evaluations is  $N \leq \frac{N^2}{P} \leq N^2$ . Specifically, we have the particular cases  $P = 1$  and  $P = N$  corresponding to the MIS schemes N3 (best performance) and N1 (worst performance), respectively. In [7], it is proved that for a specific partition with  $P$  subsets of proposals, merging any pair of subsets decreases the variance of the estimator  $\hat{I}$  of Eq. (29).

The previous idea can be applied to the other MIS schemes presented in Section V (not only N3). In particular, one can make an *a priori* partition of the proposals as in Eq. (44), and apply independently any different MIS scheme in each set. For instance, and based on some knowledge about the performance of the different proposals, one could make two disjoint sets of proposals, applying the MIS scheme N1 in the first set, and the MIS scheme N3 in the second set.

### C. Generalized Adaptive Multiple Importance Sampling

Adaptive Importance Sampling (AIS) methods update over the time the parameters of the proposal pdfs using the information of the past samples. In that way, they decrease the mismatch between the proposal and the target, and thus improve the

<sup>10</sup>Note that, for sake of simplifying the notation, we assume that all  $P$  subsets have the same number of elements. However this is not necessary, and it is straightforward to extend the conclusions of this section to the case where each subset has different number of elements

performance of the MIS scheme [8, 11, 10, 12]. The sampling and weighting options, described in this work within a static framework for sake of simplicity, can be straightforwardly applied in the adaptive context.

More specifically, let us consider a set of proposal pdfs  $\{q_{j,t}(\mathbf{x})\}$ , with  $j = 1, \dots, J$  and  $t = 1, \dots, T$ , where the subscript  $t$  indicates the iteration index of the adaptive algorithm,  $T$  is the total number of adaptation steps,  $J$  is the number of proposals per iteration, and  $N = JT$  is the total number of proposal pdfs. A general adaptation procedure takes into account, at the  $t$ -th iteration, statistical information about the target pdf gathered in all of the previous iterations,  $1, \dots, t-1$ , using one of the many algorithms that have been proposed in the literature [9, 8, 11, 10].

Hence, the sampling and the weighting procedures described in previous sections (and therefore the six MIS schemes considered in Section V) can be directly applied to the whole set of  $N$  proposal pdfs. Moreover, in the adaptive context, when many proposals are considered (the number of proposals grow over the time), the *a priori* partition approach of Section VII-B can be useful to limit the computational cost of the different MIS schemes.

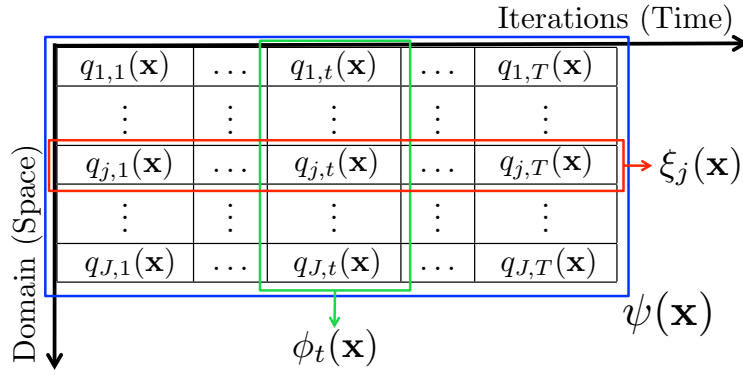


Fig. 7: Graphical representation of the  $N = JT$  proposal pdfs used in the generalized adaptive MIS scheme, spread through the state space  $\mathbb{R}^{d_x}$  ( $j = 1, \dots, J$ ) and adapted over time ( $t = 1, \dots, T$ ).

Let us assume that, at the  $t$ -th iteration, one sample is drawn from each proposal  $q_{j,t}$  (sampling  $\mathcal{S}_3$ ), i.e.,

$$\mathbf{X}_{j,t} \sim q_{j,t}(\mathbf{x}_{j,t}),$$

$j = 1, \dots, J$  and  $t = 1, \dots, T$ . Then, an importance weight  $w_{j,t}$  is assigned to each sample  $\mathbf{x}_{j,t}$ . As described exhaustively in Section IV, several strategies can be applied to build  $w_{j,t}$  considering the different MIS approaches. Fig. 7 provides a graphical representation of this scenario, by showing both the spatial and temporal evolution of the  $J = NT$  proposal pdfs. In a generic AIS algorithm, one weight

$$w_{j,t} = \frac{\pi(\mathbf{x}_{j,t})}{\varphi_{j,t}(\mathbf{x}_{j,t})}, \quad (46)$$

is associated to each sample  $\mathbf{x}_{j,t}$ . In the MIS scheme N1, the function employed in the denominator is

$$\varphi_{j,t}(\mathbf{x}) = q_{j,t}(\mathbf{x}). \quad (47)$$

In the following, we focus on the MIS scheme N3 in the adaptive framework, considering several choices of the partitioning of the set of proposals, since this scheme attains the best performance, as shown in Section 6.2. This method, with different choices of the partitioning of the set of proposals, implicitly appears in several methodologies that have been proposed independently in the literature of adaptive MIS algorithms. In the *full* N3 scheme, the function  $\varphi_{j,t}$  is

$$\varphi_{j,t}(\mathbf{x}) = \psi(\mathbf{x}) = \frac{1}{JT} \sum_{k=1}^J \sum_{r=1}^T q_{k,r}(\mathbf{x}), \quad (48)$$

where  $\psi(\mathbf{x})$  is now the mixture of all the spatial and temporal proposal pdfs. This case corresponds to the blue rectangle in Fig. 7. Furthermore, two natural alternatives of partial N3 schemes appear in this scenario. The first one uses the following partial mixture

$$\varphi_{j,t}(\mathbf{x}) = \xi_j(\mathbf{x}) = \frac{1}{T} \sum_{r=1}^T q_{j,r}(\mathbf{x}), \quad (49)$$

with  $j = 1, \dots, J$ , as mixture-proposal pdf in the IS weight denominator. Namely, in this case we consider the temporal evolution of the  $j$ -th single proposal  $q_{j,t}$ . Hence, we have  $P = J$  mixtures, each one formed by  $L = T$  components (red rectangle in Fig. 7). The other possibility is considering the mixture of all the  $q_{j,t}$ 's at the  $t$ -th iteration, i.e.,

$$\varphi_{j,t}(\mathbf{x}) = \phi_t(\mathbf{x}) = \frac{1}{J} \sum_{k=1}^J q_{k,t}(\mathbf{x}), \quad (50)$$

with  $t = 1, \dots, T$ , so that we have  $P = T$  mixtures, each one formed by  $L = J$  components (green rectangle in Fig. 7). The function  $\varphi_{j,t}$  in Eq. (47) is used in the standard PMC scheme [8]; Eq. (49), in the particular case of  $J = 1$ , has been considered in the *adaptive multiple importance sampling* (AMIS) algorithm [11]. The choice in Eq. (50) has been applied in the *adaptive population importance sampling* (APIS) algorithm [10]. In other techniques, such as Mixture PMC [21, 22, 9], a similar strategy is employed, but using sampling  $S_1$  in the mixture  $\phi_t(\mathbf{x})$ , i.e., with the MIS scheme R3.

TABLE VI: Summary of possible MIS strategies in an adaptive framework.

| MIS scheme                   | Function $\varphi_{j,t}(\mathbf{x})$  | $N$  | $P$      | $L$  | Corresponding Algorithm |
|------------------------------|---|------|----------|------|-------------------------|
|                              |   |      | $LP = N$ |      |                         |
| N1                           | $q_{j,t}(\mathbf{x})$   | $JT$ | $JT$     | 1    | PMC [8]                 |
| <i>Full</i> N3               | $\psi(\mathbf{x}) = \frac{1}{JT} \sum_{j=1}^J \sum_{t=1}^T q_{j,t}(\mathbf{x})$ |      | 1        | $JT$ | suggested in [7]        |
| <i>Partial</i> (temporal) N3 | $\xi_j(\mathbf{x}) = \frac{1}{T} \sum_{t=1}^T q_{j,t}(\mathbf{x})$              |      | $J$      | $T$  | AMIS [11], with $J = 1$ |
| <i>Partial</i> (spatial) N3  | $\phi_t(\mathbf{x}) = \frac{1}{J} \sum_{j=1}^J q_{j,t}(\mathbf{x})$             |      | $T$      | $J$  | APIS [10]               |
| <i>Partial</i> (spatial) R3  | $\phi_t(\mathbf{x}) = \frac{1}{J} \sum_{j=1}^J q_{j,t}(\mathbf{x})$             |      | $T$      | $J$  | [9, 21, 22]             |
| <i>Partial</i> (generic) N3  | generic $\varphi_{j,t}(\mathbf{x})$ in Eq. (51)                                 |      | $P$      | $L$  | suggested in [7]        |

Table VI summarizes all the possible cases discussed above. The last row corresponds to a generic grouping strategy of the proposal pdfs  $q_{j,t}$ . As previously described, we can also divide the  $N = JT$  proposals into  $P = \frac{JT}{L}$  disjoint groups of  $P$  mixtures with  $L$  components. Namely, we denote the set of  $L$  pairs of indexes corresponding to the  $p$ -th mixture ( $p = 1, \dots, P$ ) as  $\mathcal{J}_p = \{(k_{p,1}, r_{p,1}), \dots, (k_{p,L}, r_{p,L})\}$ , where  $k_{p,\ell} \in \{1, \dots, J\}$ ,  $r_{p,\ell} \in \{1, \dots, T\}$  (i.e.,  $|\mathcal{J}_p| = L$ , with each element being a pair of indexes), and  $\mathcal{J}_p \cap \mathcal{J}_q = \emptyset$  for any pair  $p, q = 1, \dots, P$ , and  $p \neq q$ . In this scenario, we have

$$\varphi_{j,t}(\mathbf{x}) = \frac{1}{L} \sum_{(k,r) \in \mathcal{J}_p} q_{k,r}(\mathbf{x}), \quad \text{with } (j,t) \in \mathcal{J}_p. \quad (51)$$

Note that using  $\psi(\mathbf{x})$  and  $\xi_j(\mathbf{x})$  the computational cost per iteration increases as the total number of iterations  $T$  grows. Indeed, at the  $t$ -th iteration all the previous proposals  $q_{j,1}, \dots, q_{j,t-1}$  (for all  $j$ ) must be evaluated at all the new samples  $\mathbf{x}_{j,t}$ . Hence, algorithms based on these proposals quickly become unfeasible as the number of iterations grows. On the other hand, using  $\phi_t(\mathbf{x})$  the computational cost per iteration is controlled by  $J$ , remaining constant regardless of the number of adaptive steps performed.

Through this subsection, we have shown that some of the most relevant adaptive MIS algorithms can be cast within the proposed generalized MIS framework. Besides this unifying perspective, new adaptive algorithms can be naturally proposed by modifying the sampling or the weighting schemes of the existing algorithms in the literature.

## VIII. NUMERICAL EXAMPLES

In the previous sections, we have provided several theoretical results for comparing different MIS schemes according to different quality measures, e.g., ranking them in terms of the variance of the corresponding estimators. In this section, we provide different numerical results in order to quantify numerically the gap among these methods. In the following, we show that even in the case where the different proposals are well tuned (in the sense of a small or no mismatch with a multimodal target), the choice of sampling and the weighing procedure dramatically affects the performance of the MIS estimator.

### A. Estimation of the normalizing constant in MIS

Let us consider a normalized bimodal target pdf  $\pi(\mathbf{x}) = \frac{1}{2}\mathcal{N}(x; \nu_1, c_1^2) + \frac{1}{2}\mathcal{N}(x; \nu_2, c_2^2)$ , with means  $\nu_1 = -3$  and  $\nu_2 = 3$ , and variances  $c_1^2 = 1$  and  $c_2^2 = 1$ . The proposal pdfs are  $q_1(x) = \mathcal{N}(x; \mu_1, \sigma_1)$  and  $q_2(x) = \mathcal{N}(x; \mu_2, \sigma_2)$ . At this point, we consider two scenarios:

- Scenario 1: In this case,  $\mu_1 = \nu_1$ ,  $\mu_2 = \nu_2$ ,  $\sigma_1^2 = c_1^2$ , and  $\sigma_2^2 = c_2^2$ . Then, both proposal pdfs can be seen as a whole mixture that exactly replicates the target, i.e.,  $\pi(\mathbf{x}) = \frac{1}{2}q_1(\mathbf{x}) + \frac{1}{2}q_2(\mathbf{x})$ . This is the desired situation pursued by an adaptive importance sampling algorithm: each proposal is centered at each target mode, and their scale parameters perfectly match the scale of the modes. Fig. 8(a) shows the target pdf in solid black line, and both proposal pdfs in blue and red dashed lines, respectively. Note that the proposals are scaled (each one integrates up to  $1/3$  so we can see the perfect matching between the target and the mixture of proposal densities).
- Scenario 2: In this case,  $\mu_1 = -2.5$ ,  $\mu_2 = 2.5$ ,  $\sigma_1^2 = 1.2$ , and  $\sigma_2^2 = 1.2$ . Therefore, there is a mismatch between the target and the two proposals. Fig. 9(a) shows the target pdf in solid black line, and both proposal pdfs in blue and red dashed lines, respectively.

The goal is estimating the normalizing constant with the six schemes described in Section V. We use the estimator of Eq. (22) with  $N = 2$  samples. We recall that depending on the sampling method, each sample might be drawn from each proposal or both samples might be drawn from the same one. In order to characterize the six estimators, we run  $2 \cdot 10^5$  simulations for each method. Note that, in this case, the true value is  $Z = 1$ .

Figure 8(b) shows a box plot representing the 25th and 75th percentiles of the distribution of the estimator  $\hat{Z}$ , obtained with different MIS methods. Each box is associated with each MIS scheme applied to the Scenario 1 described above. The blue lower and upper edges of the box correspond to the 25th and 75th percentiles, respectively, while the red line represents the median. The vertical black dashed whiskers extend to the minimum and maximum. Since the maxima cannot be appreciated in the figure in some schemes, they are displayed in Table VII jointly with the variance of  $\hat{Z}$ . Note that best methods are R3 and N3, since they always estimate perfectly the normalizing constant ( $Z = 1$ ) and the estimator has zero variance. The reason is that both methods use the whole mixture of proposals at the weight denominator, and the weights are always  $w = 1$ . Note that in this easy scenario the estimator N1, used for instance in standard PMC algorithm [8], fails dramatically. In most of the realizations  $\hat{Z}_{N1} \approx 0.5$  because each proposal (which integrates up to one) is adapted to one of the two modes (which contain roughly half of the probability mass).<sup>11</sup> Since  $E[\hat{Z}_{N1}] = Z = 1$  (as all the methods in the proposed framework), in a

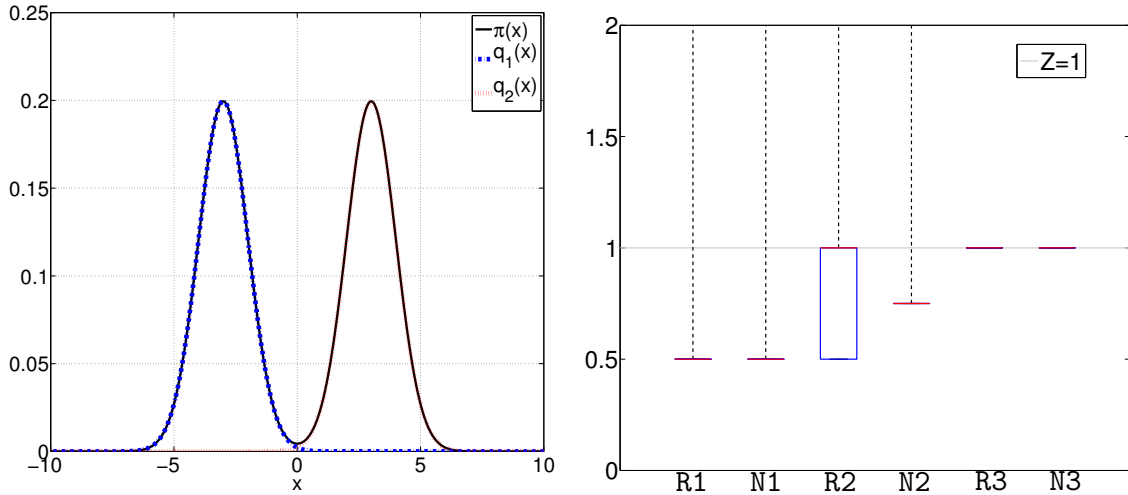
<sup>11</sup>In this setup, each proposal approximately covers a different half of the target probability mass, since each one coincides with a different mode of the target. However, in N1, the weight of each sample only accounts for its own proposal, and therefore there is not an exchange of information among the two different proposals. Note that if both proposals were covering the same mode (and therefore missing the other one), the weights would be also  $w = 0.5$  in most of the runs; the lack of information exchange between the two samples, makes impossible to know whether the target mass reported by the weight of each sample is the same and should be accounted “once”, or whether it is from another area and it should be accounted “twice”.

|                                 | MIS Scheme            | R1    | N1     | R2     | N2     | R3   | N3   |
|---------------------------------|-----------------------|-------|--------|--------|--------|------|------|
| <b>Perfect matching (Sc. 1)</b> | Max. value            | 7298  | 35864  | 51395  | 260525 | 1    | 1    |
|                                 | $\text{Var}(\hat{Z})$ | 74925 | 7891.2 | 46.2   | 2284.6 | 0    | 0    |
| <b>Mismatch (Sc. 2)</b>         | Max. value            | 39250 | 77238  | 132556 | 97056  | 1.59 | 1.59 |
|                                 | $\text{Var}(\hat{Z})$ | 1847  | 6874   | 10285  | 5474   | 0.01 | 0.01 |

TABLE VII: (Ex. of Section. VIII-A) Maximum value of the estimator  $\hat{Z}$  in  $2 \cdot 10^5$  runs for each MIS scheme, in two different scenarios.

few runs the value  $\hat{Z}_{N1}$  is extremely high as shown in VII). This huge values occur when a sample drawn from the tail of the proposal falls close to the other mode of the target (where actually the other proposal is placed). This simple example shows the variance reduction that can be attained in MIS schemes by using all the proposals at the weight denominator.

Figure 9(b) shows an equivalent box plot for Scenario 2. In this case, the mismatch between proposals and target pdfs worsens all MIS schemes. Note that R3 and N3 do not perfectly approximate  $Z$ , but still largely outperform the other schemes. In particular, the median is still around  $\hat{Z} = 1$  and their variance is smaller. The maximum values and the variance of  $\hat{Z}$  in all methods are again provided in Table VII.



(a) Distributions in the perfect matching case (Scenario 1).

(b) Boxplot showing the 25th and 75th percentiles of the distribution of the estimator  $\hat{Z}$  (where  $Z = 1$ ), obtained with different MIS schemes in the perfect matching case (Scenario 1). The red lines represent the medians of these distributions. The maximum values of R1, N1, R2 and N2 are not depicted in the figure since they exceed the limits (the values are in Table VII). The boxes of S3 and N3 are collapsed in a line at 1 (maximum, minimum, and median coincide).

Fig. 8: (Ex. of Section VIII-A) Scenario 1 (perfect matching). Estimation of the normalizing constant.

#### B. Estimation of an expected value in a perfect proposal-target matching scenario

Let us consider as a target pdf a mixture of three Gaussian pdfs,

$$\pi(\mathbf{x}) = \frac{1}{3}\mathcal{N}(x; \nu_1, c_1) + \frac{1}{3}\mathcal{N}(x; \nu_2, c_2) + \frac{1}{3}\mathcal{N}(x; \nu_3, c_3),$$

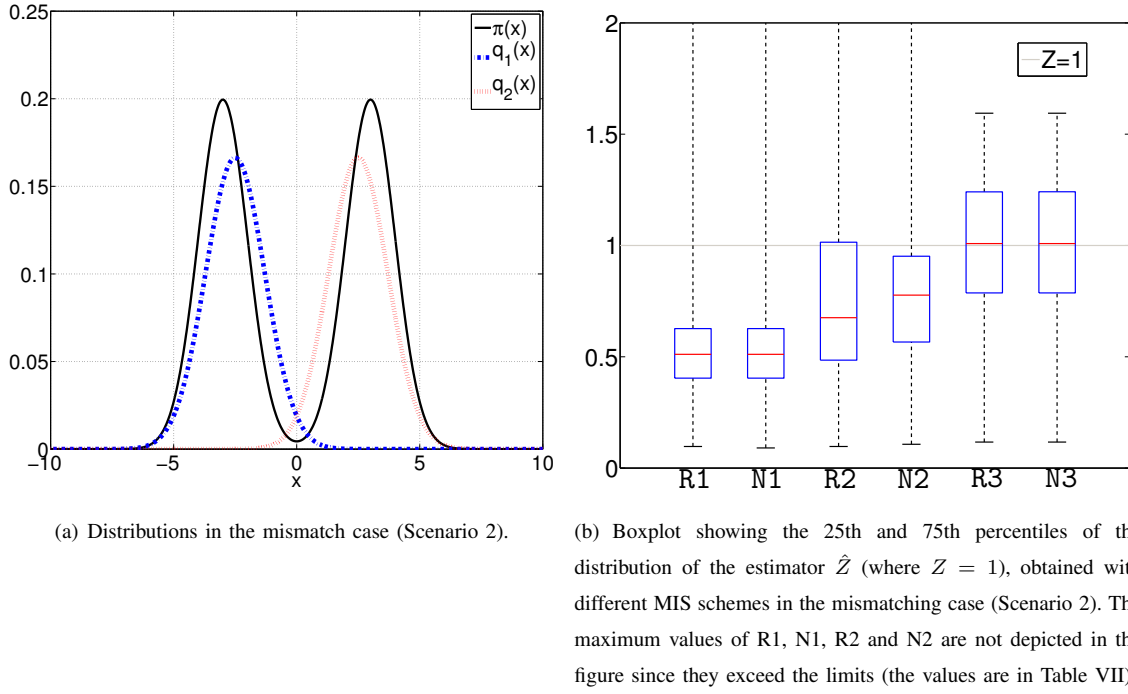


Fig. 9: (Ex. of Section. VIII-A) Scenario 2 (proposals-target mismatch). Estimation of the normalizing constant.

with means  $\nu_1 = -3$ ,  $\nu_2 = 0$ , and  $\nu_3 = 3$ , and variances  $c_1 = 1, c_2 = 1$ , and  $c_3 = 1$ . As proposal functions we use  $q_i(x) = \mathcal{N}(x; \mu_i, \sigma_i)$ , with  $\mu_i = \nu_i$ , and  $\sigma_i = c_i$  with  $i = 1, 2, 3$ , i.e., the proposal pdfs can be seen as a whole mixture that exactly replicates the target, i.e.,  $\pi(x) = \psi(x) = \frac{1}{3}q_1(x) + \frac{1}{3}q_2(x) + \frac{1}{3}q_3(x)$ . All the pdfs are represented in Fig. 10. Note that the proposals are scaled (they integrate up to  $1/3$ ), so it can be displayed the perfect match between mixture of proposals and the target.

The goal is estimating the mean of the target pdf with the six MIS schemes. Fig. 11 shows the MSE of the estimator  $\hat{I}$  for all the methods w.r.t. the number of total samples (note that some schemes require that the total number of samples is multiple of  $M = 3$ ). The results have been averaged over  $5 \cdot 10^6$  runs. The solid black line shows the variance of the estimator  $\bar{I}$  of Eq. (37), i.e. sampling directly from the target pdf (since this is possible in this easy example). Note that the method  $\hat{I}_{R3}$  exactly replicates the performance of  $\bar{I}$ : this method samples from the mixture of Gaussians in the traditional way and the weights, due to the perfect match, are always  $w = 1$ , i.e.,  $\hat{I}_{R3}$  and  $\bar{I}$  are equivalent. We can see that  $\hat{I}_{N3}$  is the best estimator in terms of variance, while  $\hat{I}_{R1}$  and  $\hat{I}_{N1}$  present a high variance. Note that, surprisingly,  $\hat{I}_{N3}$  has better performance than sampling from the target, i.e., estimator  $\bar{I}$ . This is because the sampling  $S_3$  can be seen as a sampling from the mixture of proposals  $\psi(x)$  (which coincides with the target in this example) with a variance reduction technique, as we discuss in Appendix A. Note also that the inequality proved in Theorem 1 holds since all methods are unbiased and therefore the MSE is due only to the variance. We can see that  $\hat{I}_{R2}$  and  $\hat{I}_{N2}$  behave also bad in terms of variance.

Figure 12 shows the variance of the estimator  $\tilde{I}$  of Eq. (28) for all methods. First, note that the MSE of R3 and N3 is the same than in Fig. 11, since the estimators  $\hat{I}$  and  $\tilde{I}$  are equivalent in this scenario (since they perfectly estimate the normalizing constant, i.e.,  $\hat{Z} = Z$ ). Note that the relations observed and proved for the different MIS schemes in terms of the variance of the estimator  $\hat{I}$ , are also kept here when we increase the number of samples.

Figure 13 shows the averaged Effective Sample Size (ESS) calculated with the approximation of Eq. (40). We can see that in this perfect matching scenario, R3 and N3 always obtain exactly  $\widehat{ESS} = M$  since weights have the same value, even though



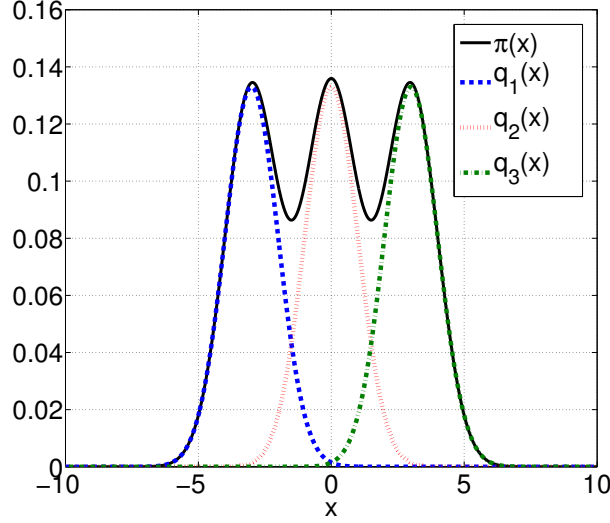


Fig. 10: (Ex. of Section. VIII-B) Target and proposal pdfs.

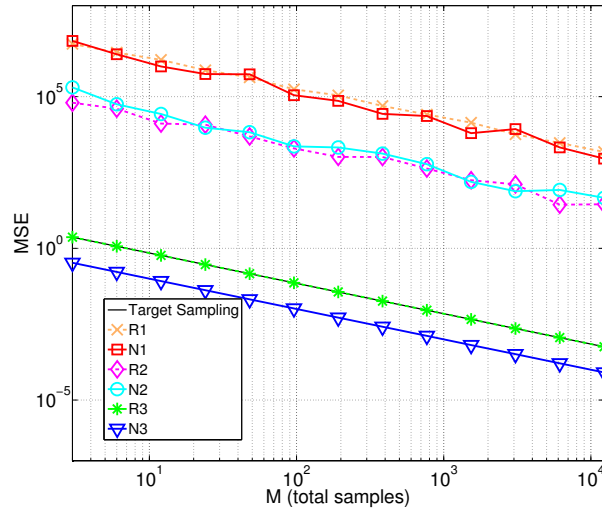


Fig. 11: (Ex. of Section. VIII-B) MSE of the MIS estimator  $\hat{I}$  (unnormalized weights) for the different MIS schemes.

N3 outperforms R3 in terms of variance. Observing the other MIS schemes, it seems that the ranking in terms of variances is also preserved in terms of  $\widehat{ESS}$ .

### C. Multidimensional mixture of generalized Gaussian distributions

Let us consider a mixture of multivariate generalized Gaussian distributions (GGD) as a target pdf. In particular

$$\pi(\mathbf{x}) = \frac{1}{3} \sum_{k=1}^3 \mathcal{GG}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\alpha}_k, \boldsymbol{\beta}_k), \quad \mathbf{x} \in \mathbb{R}^{D_x}, \quad (52)$$

where  $\boldsymbol{\mu}_k = [\mu_{k,1}, \dots, \mu_{k,D_x}]^\top$ ,  $\boldsymbol{\alpha}_k = [\alpha_{k,1}, \dots, \alpha_{k,D_x}]^\top$ , and  $\boldsymbol{\beta}_k = [\beta_{k,1}, \dots, \beta_{k,D_x}]^\top$  are respectively the mean, scale, and shape parameters of each component of the mixture. Each component of the mixture factorizes in all dimensions, i.e., the

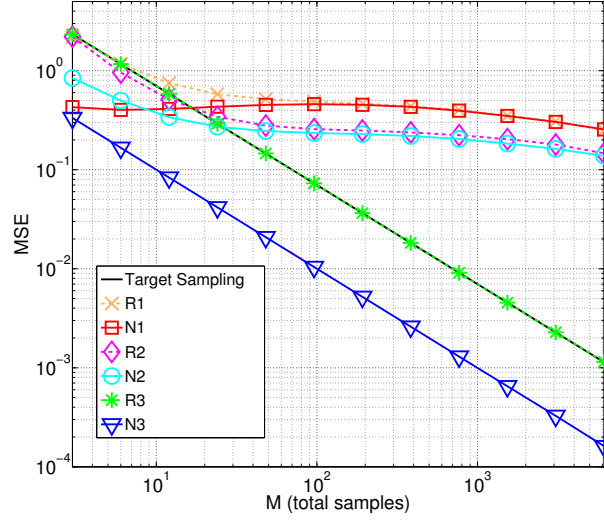


Fig. 12: **(Ex. of Section. VIII-B)** MSE of the MIS estimator  $\tilde{I}$  (normalized weights) for the different MIS schemes.

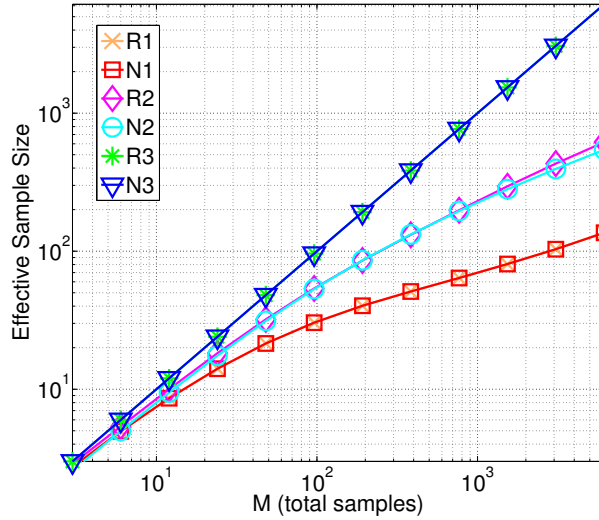


Fig. 13: **(Ex. of Section. VIII-B)** Effective Sample Size ( $\widehat{ESS}$ ) calculated as in Eq. (40) for the different MIS schemes.

multivariate GGD pdf is the product of  $N$  unidimensional GGD pdfs. Namely,

$$\mathcal{GG}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\alpha}_k, \boldsymbol{\beta}_k) = \prod_{d=1}^{D_x} \kappa_{k,d} \exp \left( - \left( \frac{|x_d - \mu_{k,d}|}{\alpha_{k,d}} \right)^{\beta_{k,d}} \right),$$

where  $\kappa_{k,d} = \frac{\beta_{k,d}}{2\alpha_{k,d}\Gamma(\frac{1}{\beta_{k,d}})}$ ,  $\Gamma(\cdot)$  is the gamma function, and  $x_d$  is the  $d$ -th dimension of  $\mathbf{x}$ . This family of distributions includes both Gaussian and Laplace distributions with  $\beta = 2$  and  $\beta = 1$ , respectively.

In this example,  $\mu_{1,d} = -3$ ,  $\mu_{2,d} = 1$ ,  $\mu_{3,d} = 5$ ,  $\beta_{1,d} = 1.1$ ,  $\beta_{2,d} = 1.8$ ,  $\beta_{3,d} = 5$ ,  $\alpha_{1,d} = \alpha_{2,d} = \alpha_{3,d} = 1$  for all  $d = 1, \dots, D_x$ . The expected value of the target  $\pi(\mathbf{x})$  is then  $E_\pi[X_d] = 1$  for  $d = 1, \dots, D_x$ . In order to study the performance of the different MIS schemes, we vary the dimension of the state space in Eq. (52) testing different values of  $D_x$  (with  $2 \leq D_x \leq 10$ ).

We consider the problem of approximating via Monte Carlo the expected value of the target density, and we compare the

performance of all MIS schemes. In this example, we use  $N = 500$  non-standardized t-student densities as proposal functions, where each location parameter has been selected uniformly within the  $[-6, 6]^{D_x}$  square, and the scale parameters and the degree of freedom parameters have been selected as  $\sigma_{n,d} = 5$  and  $\nu_{n,d} = 5$ , respectively, for  $n = 1, \dots, N$  and  $d = 1, \dots, D_x$ . For each method, we draw  $M = kN$  samples, with  $k = 32$ , and we average all the results over 200 runs.

Fig. 14 shows the MSE in the estimation of the mean of the target (averaged over all dimensions) when we increase the dimension  $D_x$ . Note that the hierarchy established in Section VI-A also holds in this example regardless the dimension. In this case, methods R1 and N1 behave poorly even at lower dimensions, while the other MIS schemes have a similar behavior. When we increase the dimension, all the methods degrade, and, at certain point ( $D_x \geq 6$ ), the performance of all of them is similar. Note that in this example, the proposal pdfs are fixed in random locations of the space, which can be considered covered at low dimensions (since we are using  $N = 500$  pdfs), but this coverage becomes worse as the dimension increases. This can probably explain the similar performance of all the methods in higher dimensions. If we performed adaptive MIS algorithms in order to adapt the proposal pdfs, we would expect that the MIS scheme N3 outperformed the other methods substantially as in previous examples.

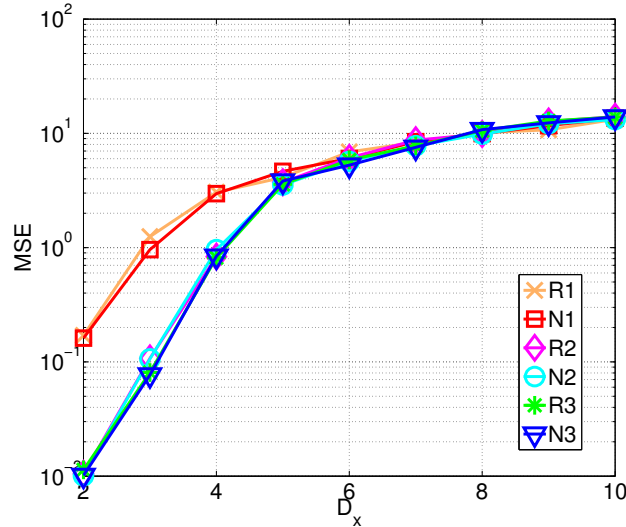


Fig. 14: (Ex. of Section. VIII-C) MSE of the MIS estimator  $\tilde{I}$  (normalized weights) for the different MIS schemes when we increase the dimension  $D_x$  of the state space.

#### D. Discussion on the experimental results

First of all, note that the numerical simulations provided in this section corroborate the theoretical results of Section VI. More specifically, the hierarchy shown in Fig. 11, based on MSE of  $\hat{I}$ , corresponds to the hierarchy in terms of variance of  $\hat{I}$  given in Theorems 1 and 2 (the latter for the case of  $N = 2$  proposals). The same hierarchy is represented graphically in Fig. 14 and, in terms of  $\widehat{ESS}$ , in Fig. 13. Furthermore, Fig. 12 depicts the MSE of  $\tilde{I}$ : for large enough values of  $M$  (so that a good approximation of  $Z$  is attained), the MIS schemes are ordered exactly as in Fig. 11 (as discussed in Section VI-A).

The numerical experiments confirm that N3 provides the best performance. The scheme R3 also presents a good performance in most cases. A possible interpretation is the following: N3 and R3 apply the whole mixture at the denominator of each weight, thus providing an exchange of information between all the proposals. This exchange of information is essential in multimodal

scenarios, where the whole set of proposals, seen as a mixture, should mimic the whole target, but each proposal should adapt locally to the target. Since the variance of the IS weight depends on the mismatch of the target (numerator) w.r.t. proposal (denominator), the use of the whole mixture in the denominator reduces the variance of the weight in general, and therefore, also the variance of the estimator (see the variance analysis in Appendix A). The scheme N3 goes a step further w.r.t. R3, drawing deterministically one sample from each mixand of  $\psi(\mathbf{x})$ , which can be seen as drawing  $N$  samples from the mixture  $\psi(\mathbf{x})$  with a modified version of stratified sampling, a well-known variance reduction technique (see Appendix A and [3, Section 9.12]).

The performance of R1 and N1 is, in general, worse than the performance of the other schemes. Both schemes account at the weight denominator only for the proposal from which the sample is drawn, which in a multimodal scenario can be problematic (see the example of Section VIII-A). While R1 is a novel scheme that has naturally arisen in this work, and it probably has little interest from a practical point of view, N1 has been applied in different adaptive MIS algorithms, such as the original version of PMC [8].

The schemes R2 and N2, that have appeared in this new framework, probably deserve a further analysis. The hierarchy theoretically proved for  $N = 2$  proposals in Theorem 2 still holds in the numerical examples for  $N > 2$ , e.g. in Figs. 11 and 12. In some scenarios, for instance where there is a big number of proposals compared to the modes of the target, these schemes can attain most of the variance reduction of N1 and N3 while reducing the number of proposal evaluations.

Finally, observe that in Fig. 12, when a small number of samples  $M$  is employed, the schemes N1, N2 and N3, i.e., those with index selection without replacement ( $\mathcal{S}_2$  and  $\mathcal{S}_3$ ), behave better. This occurs because, in this case, the variance associated to the index selection is reduced by guaranteeing that all proposal pdfs are always used.

## IX. CONCLUSIONS

In this work, we have introduced a unified framework for sampling and weighting in the context of multiple importance sampling (MIS). This framework extends the concept of a proper weighted sample, enabling the design of a wide range of sampling/weighting combinations. In particular, we have considered three specific sampling procedures and we have proposed five types of generic weighting functions (related to different conditional and marginal distributions which depend on the sampling scheme). Moreover, as a result of the combinations of sampling and weighting procedures, we have analyzed six valid schemes combining adequately the sampling procedures with the corresponding proper weighting functions (three of them are not present in the literature to the best of our knowledge). Moreover, we have provided a theoretical comparison of these schemes in terms of variance and effective sample size, establishing a ranking of the different methods in terms of performance. We have analyzed the behavior of these methods in three different numerical examples which corroborate the previous theoretical analysis. Furthermore, the computational cost and the application of the proposed schemes to adaptive importance sampling methods have also been discussed, describing different possible efficient MIS schemes when a population of proposals is adapted online. Finally, a novel effective sample size measure for MIS has been proposed, and the multinomial resampling procedure has been analyzed for the different MIS schemes.

## REFERENCES

- [1] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, 2004.
- [2] J. S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer, 2004.
- [3] A. Owen. *Monte Carlo theory, methods and examples*. <http://statweb.stanford.edu/~owen/mc/>, 2013.

- [4] F. Liang. Dynamically weighted importance sampling in Monte Carlo computation. *Journal of the American Statistical Association*, 97(459):807–821, 2002.
- [5] E. Veach and L. Guibas. Optimally combining sampling techniques for Monte Carlo rendering. In *SIGGRAPH 1995 Proceedings*, pages 419–428, 1995.
- [6] A. Owen and Y. Zhou. Safe and effective importance sampling. *Journal of the American Statistical Association*, 95(449):135–143, 2000.
- [7] V. Elvira, L. Martino, D. Luengo, and M. F. Bugallo. Efficient multiple importance sampling estimators. *Signal Processing Letters, IEEE*, 22(10):1757–1761, 2015.
- [8] O. Cappé, A. Guillin, J. M. Marin, and C. P. Robert. Population Monte Carlo. *Journal of Computational and Graphical Statistics*, 13(4):907–929, 2004.
- [9] O. Cappé, R. Douc, A. Guillin, J. M. Marin, and C. P. Robert. Adaptive importance sampling in general mixture classes. *Statistics and Computing*, 18:447–459, 2008.
- [10] L. Martino, V. Elvira, D. Luengo, and J. Corander. An adaptive population importance sampler: Learning from the uncertainty. *IEEE Transactions on Signal Processing*, 63(16):4422–4437, 2015.
- [11] J. M. Cornuet, J. M. Marin, A. Mira, and C. P. Robert. Adaptive multiple importance sampling. *Scandinavian Journal of Statistics*, 39(4):798–812, December 2012.
- [12] M. F. Bugallo, L. Martino, and J. Corander. Adaptive importance sampling in signal processing. (*To appear*) *Digital Signal Processing*, 2015.
- [13] N. Gordon, D. Salmond, and A. F. M. Smith. Novel approach to nonlinear and non-Gaussian Bayesian state estimation. *IEE Proceedings-F Radar and Signal Processing*, 140:107–113, 1993.
- [14] Herman Kahn and Andy W Marshall. Methods of reducing sample size in monte carlo computations. *Journal of the Operations Research Society of America*, 1(5):263–278, 1953.
- [15] J. Geweke. Bayesian inference in econometric models using Monte Carlo integration. *Econometrica*, 24:1317–1399, 1989.
- [16] A. Kong. A note on importance sampling using standardized weights. *University of Chicago, Dept. of Statistics, Tech. Rep*, 348, 1992.
- [17] A. Kong, J. S. Liu, and W. H. Wong. Sequential imputations and Bayesian missing data problems. *Journal of the American Statistical Association*, 9:278–288, 1994.
- [18] A. Doucet, S. Godsill, and C. Andrieu. On sequential Monte Carlo Sampling methods for Bayesian filtering. *Statistics and Computing*, 10(3):197–208, 2000.
- [19] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Klapp. A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Transactions Signal Processing*, 50(2):174–188, February 2002.
- [20] A. Doucet, N. de Freitas, and N. Gordon, editors. *Sequential Monte Carlo Methods in Practice*. Springer, New York, 2001.
- [21] G.R. Douc, J.M. Marin, and C. Robert. Convergence of adaptive mixtures of importance sampling schemes. *Annals of Statistics*, 35:420–448, 2007.
- [22] G.R. Douc, J.M. Marin, and C. Robert. Minimum variance importance sampling via population Monte Carlo. *ESAIM: Probability and Statistics*, 11:427–447, 2007.
- [23] H. Niederreiter. *Random Number Generation and Quasi-Monte Carlo Methods*. Society for Industrial Mathematics, 1992.

- [24] G. H. Hardy, J. E. Littlewood, and G. Pólya. *Inequalities*. Cambridge Univ. Press, 1952.
- [25] M. Abramowitz and I. A. Stegun. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*. Dover Pub., number 55., 1972.
- [26] P. W. Gwanyama. The HM-GM-AM-QM inequalities. *The College Mathematics Journal*, 35(1):47–50, Jan. 2004.

## APPENDIX

Let us recall again the sampling procedure  $\mathcal{S}_3$  (i.e., with deterministic selection of the index):

1) For  $n = 1, \dots, N$  :

a) Draw the sample  $\mathbf{x}_n \sim q_n(\mathbf{x})$ .

In this case, obviously  $\mathbf{x}_n \sim q_n(\mathbf{x})$  for  $n = 1, \dots, N$ . However, if the samples  $\mathbf{x}_1, \dots, \mathbf{x}_N$  are used jointly regardless of their order, then it can be interpreted that all of them have been drawn from the following mixture (see [3, Section 9.12]):

$$\psi(\mathbf{x}) = \frac{1}{N} \sum_{n=1}^N q_n(\mathbf{x}). \quad (53)$$

More formally, a sample  $\mathbf{z}$  chosen uniformly within the set  $\{\mathbf{x}_n\}_{n=1}^N$ , is distributed as  $\psi(\mathbf{x})$ . Namely, drawing an index  $k^*$  uniformly on  $\{1, \dots, N\}$ , then we can write

$$\mathbf{z} = \mathbf{x}_{k^*} \sim \psi(\mathbf{x}).$$

An equivalent interpretation can be obtained considering a random permutation of the set  $\{\mathbf{x}_n\}_{n=1}^N$ , denoted as

$$\{\mathbf{z}_1 = \mathbf{x}_{k_1}, \dots, \mathbf{z}_N = \mathbf{x}_{k_N}\},$$

where  $\cup_{n=1}^N k_n = \{1, \dots, N\}$ . Then, we have  $\mathbf{z}_n \sim \psi(\mathbf{x})$  for  $n = 1, \dots, N$ . The procedure  $\mathcal{S}_3$  follows similar principle as a well-known variance reduction method, known as the stratified sampling [1, 2], where the domain of  $\mathbf{x}$  is divided into different regions that, in this case, are unbounded and overlapped [3, Section 9.12]. Finally, note that the approach  $\mathcal{S}_3$  can also be seen as the application of a quasi-Monte Carlo technique [23] for generating the deterministic sequence of indexes  $k_1 = 1, k_2 = 2, \dots, k_N = N$  (uniform, in the sense of low-discrepancy sequence [23]) and then drawing  $\mathbf{x}_{k_n} \sim q_{k_n}(\mathbf{x})$  for  $n = 1, \dots, N$ .

In this appendix we prove the unbiasedness of the estimator  $\hat{I}$  of Eq. (21) for the five weighting options described in Section IV. We recall that the general expression for the expectation of  $\hat{I}$  within the proposed framework is

$$E[\hat{I}] = \frac{1}{ZN} \sum_{n=1}^N \sum_{j_{1:N}} \int \frac{\pi(\mathbf{x}_n)g(\mathbf{x}_n)}{\varphi_{\mathcal{P}_n}(\mathbf{x}_n)} P(j_{1:N}) p(\mathbf{x}_n | j_n) d\mathbf{x}_n. \quad (54)$$

**OPTION 1** ( $\mathcal{W}_1$ ):  $\varphi_{\mathcal{P}_n}(\mathbf{x}_n) = \varphi_{j_{1:n-1}}(\mathbf{x}_n) = p(\mathbf{x}_n | j_{1:n-1})$ . We first marginalize in Eq. (54) over all indexes that do not affect the  $n$ -th weight:

$$\begin{aligned} E[\hat{I}] &= \frac{1}{ZN} \sum_{n=1}^N \sum_{j_{1:n-1}} \sum_{j_{n:N}} \int \frac{\pi(\mathbf{x}_n)g(\mathbf{x}_n)}{\varphi_{j_{1:n-1}}(\mathbf{x}_n)} p(\mathbf{x}_n, j_{1:N}) d\mathbf{x}_n \\ &= \frac{1}{ZN} \sum_{n=1}^N \sum_{j_{1:n-1}} \int \frac{\pi(\mathbf{x}_n)g(\mathbf{x}_n)}{\varphi_{j_{1:n-1}}(\mathbf{x}_n)} \sum_{j_{n:N}} p(\mathbf{x}_n, j_{1:N}) d\mathbf{x}_n \\ &= \frac{1}{ZN} \sum_{n=1}^N \sum_{j_{1:n-1}} \int \frac{\pi(\mathbf{x}_n)g(\mathbf{x}_n)}{\varphi_{j_{1:n-1}}(\mathbf{x}_n)} p(\mathbf{x}_n, j_{1:n-1}) d\mathbf{x}_n \\ &= \frac{1}{ZN} \sum_{n=1}^N \sum_{j_{1:n-1}} \int \frac{\pi(\mathbf{x}_n)g(\mathbf{x}_n)}{\varphi_{j_{1:n-1}}(\mathbf{x}_n)} p(\mathbf{x}_n | j_{1:n-1}) P(j_{1:n-1}) d\mathbf{x}_n \end{aligned} \quad (55)$$

Then, substituting  $\varphi_{j_{1:n-1}}(\mathbf{x}_n) = p(\mathbf{x}_n|j_{1:n-1})$  into Eq. (55), we have:

$$\begin{aligned} E[\hat{I}] &= \frac{1}{ZN} \sum_{n=1}^N \sum_{j_{1:n-1}} \int \pi(\mathbf{x}_n) g(\mathbf{x}_n) P(j_{1:n-1}) d\mathbf{x}_n \\ &= \frac{1}{ZN} \sum_{n=1}^N \int \pi(\mathbf{x}_n) g(\mathbf{x}_n) d\mathbf{x}_n \\ &= \frac{1}{Z} \int \pi(\mathbf{x}) g(\mathbf{x}) d\mathbf{x} = I. \quad \square \end{aligned}$$

**OPTION 2** ( $\mathcal{W}_2$ ):  $\varphi_{\mathcal{P}_n}(\mathbf{x}_n) = \varphi_{j_n}(\mathbf{x}_n) = p(\mathbf{x}_n|j_n)$ . We substitute  $\varphi_{j_n}(\mathbf{x}_n) = p(\mathbf{x}_n|j_n)$ , into Eq. (54):

$$\begin{aligned} E[\hat{I}] &= \frac{1}{ZN} \sum_{n=1}^N \sum_{j_{1:N}} \int \frac{\pi(\mathbf{x}_n) g(\mathbf{x}_n)}{p(\mathbf{x}_n|j_n)} P(j_{1:N}) p(\mathbf{x}_n|j_n) d\mathbf{x}_n \\ &= \frac{1}{ZN} \sum_{n=1}^N \sum_{j_{1:N}} \int \pi(\mathbf{x}_n) g(\mathbf{x}_n) P(j_{1:N}) d\mathbf{x}_n \\ &= \frac{1}{ZN} \sum_{n=1}^N \int \pi(\mathbf{x}_n) g(\mathbf{x}_n) d\mathbf{x}_n \\ &= \frac{1}{Z} \int \pi(\mathbf{x}) g(\mathbf{x}) d\mathbf{x} = I. \quad \square \end{aligned}$$

**OPTION 3** ( $\mathcal{W}_3$ ):  $\varphi_{\mathcal{P}_n}(\mathbf{x}_n) = \varphi_n(\mathbf{x}_n) = p(\mathbf{x}_n)$ . Since  $\varphi_n$  does not depend on any index, we can first marginalize over the whole set of indexes  $j_{1:N}$  in Eq. (54):

$$\begin{aligned} E[\hat{I}] &= \frac{1}{ZN} \sum_{n=1}^N \int \frac{\pi(\mathbf{x}_n) g(\mathbf{x}_n)}{\varphi_n(\mathbf{x}_n)} \left[ \sum_{j_{1:N}} p(\mathbf{x}_n, j_{1:N}) \right] d\mathbf{x}_n \\ &= \frac{1}{ZN} \sum_{n=1}^N \int \frac{\pi(\mathbf{x}_n) g(\mathbf{x}_n)}{\varphi_n(\mathbf{x}_n)} p(\mathbf{x}_n) d\mathbf{x}_n. \end{aligned} \quad (56)$$

Then, substituting  $\varphi_n = p(\mathbf{x}_n)$  in Eq. (56):

$$\begin{aligned} E[\hat{I}] &= \frac{1}{ZN} \sum_{n=1}^N \int \pi(\mathbf{x}_n) g(\mathbf{x}_n) d\mathbf{x}_n \\ &= \frac{1}{Z} \int \pi(\mathbf{x}) g(\mathbf{x}) d\mathbf{x} = I. \quad \square \end{aligned}$$

**OPTION 4** ( $\mathcal{W}_4$ ):  $\varphi_{\mathcal{P}_n}(\mathbf{x}) = \varphi_{j_{1:N}}(\mathbf{x}) = f(\mathbf{x}|j_{1:N}) = \frac{1}{N} \sum_{n=1}^N q_{j_n}(\mathbf{x})$ . In this case, the expectation of  $\hat{I}$  can be expressed as:

$$\begin{aligned} E[\hat{I}] &= \frac{1}{ZN} \sum_{j_{1:N}} P(j_{1:N}) \int \sum_{n=1}^N \frac{\pi(\mathbf{x}_n) g(\mathbf{x}_n)}{\varphi_{j_{1:N}}(\mathbf{x}_n)} q_{j_n}(\mathbf{x}_n) d\mathbf{x}_n \\ &= \frac{1}{ZN} \sum_{j_{1:N}} P(j_{1:N}) \int \frac{\pi(\mathbf{x}) g(\mathbf{x})}{\varphi_{j_{1:N}}(\mathbf{x})} \sum_{n=1}^N q_{j_n}(\mathbf{x}) d\mathbf{x}. \end{aligned} \quad (57)$$

Substituting  $\varphi_{j_{1:N}}(\mathbf{x}) = f(\mathbf{x}|j_{1:N}) = \frac{1}{N} \sum_{n=1}^N q_{j_n}(\mathbf{x})$  in Eq. (57):

$$\begin{aligned} E[\hat{I}] &= \frac{1}{Z} \sum_{j_{1:N}} \int \frac{\pi(\mathbf{x}) g(\mathbf{x})}{\frac{1}{N} \sum_{n=1}^N q_{j_n}(\mathbf{x})} \left[ \frac{1}{N} \sum_{n=1}^N q_{j_n}(\mathbf{x}) \right] P(j_{1:N}) d\mathbf{x} \\ &= \frac{1}{Z} \sum_{j_{1:N}} \int \pi(\mathbf{x}) g(\mathbf{x}) P(j_{1:N}) d\mathbf{x} \\ &= \frac{1}{Z} \int \pi(\mathbf{x}) g(\mathbf{x}) d\mathbf{x} = I. \quad \square \end{aligned}$$

**OPTION 5** ( $\mathcal{W}_5$ ):  $\varphi_{\mathcal{P}_n}(\mathbf{x}) = \varphi(\mathbf{x}) = f(\mathbf{x}) = \frac{1}{N} \sum_{n=1}^N q_n(\mathbf{x}) = \psi(\mathbf{x})$ . Now, the expectation of  $\hat{I}$  becomes

$$\begin{aligned} E[\hat{I}] &= \frac{1}{ZN} \sum_{j_{1:N}} P(j_{1:N}) \int \sum_{n=1}^N \frac{\pi(\mathbf{x}_n)g(\mathbf{x}_n)}{\varphi(\mathbf{x}_n)} q_{j_n}(\mathbf{x}_n) d\mathbf{x}_n \\ &= \frac{1}{Z} \int \frac{\pi(\mathbf{x})g(\mathbf{x})}{\varphi(\mathbf{x})} \sum_{j_{1:N}} \left[ \frac{1}{N} \sum_{n=1}^N q_{j_n}(\mathbf{x}) \right] P(j_{1:N}) d\mathbf{x} \\ &= \frac{1}{Z} \int \frac{\pi(\mathbf{x})g(\mathbf{x})}{\varphi(\mathbf{x})} \psi(\mathbf{x}) d\mathbf{x}, \end{aligned} \quad (58)$$

where, in the last step, we have used the fact that

$$\sum_{j_{1:N}} \left[ \frac{1}{N} \sum_{n=1}^N q_{j_n}(\mathbf{x}) \right] P(j_{1:N}) = f(\mathbf{x}) = \psi(\mathbf{x})$$

for any valid sampling procedure within this framework (see Remark 1 and Section III-D for more details). Substituting  $\varphi(\mathbf{x}) = \psi(\mathbf{x})$  in Eq. (58)

$$\begin{aligned} E[\hat{I}] &= \frac{1}{Z} \int \frac{\pi(\mathbf{x})g(\mathbf{x})}{\psi(\mathbf{x})} \psi(\mathbf{x}) d\mathbf{x} \\ &= \frac{1}{Z} \int \pi(\mathbf{x})g(\mathbf{x}) d\mathbf{x} = I. \quad \square \end{aligned} \quad (59)$$

Let us consider the unbiased estimator,

$$\hat{I} = \frac{1}{ZN} \sum_{n=1}^N w_n(\mathbf{x}_n)g(\mathbf{x}_n), \quad (60)$$

that approximates  $I$ . Then, the variance of  $\hat{I}$  can be expressed in the general form as

$$\begin{aligned} \text{Var}(\hat{I}) &= E_{p(\mathbf{x}_{1:N}, j_{1:N})} \left[ \left( \hat{I} - E_{p(\mathbf{x}_{1:N}, j_{1:N})}[\hat{I}] \right)^2 \right] \\ &= E_{p(\mathbf{x}_{1:N}, j_{1:N})}[\hat{I}^2] - E_{p(\mathbf{x}_{1:N}, j_{1:N})}[\hat{I}]^2. \end{aligned} \quad (61)$$

In the general case of Eq. (61), the  $N$  terms of the sum of the estimator in  $\hat{I}$  are dependent. However, in the specific cases where they are independent, the variance of a sum of r.v.'s can be simplified as the sum of the variances, i.e.,

$$\begin{aligned} \text{Var}(\hat{I}) &= \frac{1}{Z^2 N^2} \left[ \sum_{n=1}^N E_{p(\mathbf{x}_n, j_n)}[w_n^2(\mathbf{x}_n)g^2(\mathbf{x}_n)] \right. \\ &\quad \left. - \sum_{n=1}^N E_{p(\mathbf{x}_n, j_n)}^2[w_n(\mathbf{x}_n)g(\mathbf{x}_n)] \right] \\ &= \frac{1}{Z^2 N^2} \left[ \sum_{n=1}^N \sum_{j_n=1}^N \int \frac{\pi^2(\mathbf{x}_n)g^2(\mathbf{x}_n)}{\varphi_{\mathcal{P}_n}^2(\mathbf{x}_n)} p(\mathbf{x}_n|j_n) P(j_n) d\mathbf{x}_n \right. \\ &\quad \left. - \sum_{n=1}^N \left( \sum_{j_n=1}^N \int \frac{\pi(\mathbf{x}_n)g(\mathbf{x}_n)}{\varphi_{\mathcal{P}_n}(\mathbf{x}_n)} p(\mathbf{x}_n|j_n) P(j_n) d\mathbf{x}_n \right)^2 \right]. \end{aligned} \quad (62)$$

In some MIS schemes, the  $N$  terms are dependent (due to a sampling without replacement or because the  $n$ -th weight depends on several indexes  $j_k$ , with at least one  $k \neq n$ ). Nevertheless, conditioned to the whole set of indexes  $j_{1:N}$ , the terms



of the sum in Eq. (60) are always conditionally independent, so we can apply

$$\begin{aligned}
\text{Var}(\hat{I}) &= \frac{1}{Z^2 N^2} \sum_{j_{1:N}} \left[ \sum_{n=1}^N E_{p(\mathbf{x}_n|j_n)} [w_n^2(\mathbf{x}_n) g^2(\mathbf{x}_n)] \right. \\
&\quad \left. - \sum_{n=1}^N E_{p(\mathbf{x}_n|j_n)}^2 [w_n(\mathbf{x}_n) g(\mathbf{x}_n)] \right] P(j_{1:N}) \\
&= \frac{1}{Z^2 N^2} \sum_{j_{1:N}} \left[ \sum_{n=1}^N \int \frac{\pi^2(\mathbf{x}_n) g^2(\mathbf{x}_n)}{\varphi_{\mathcal{P}_n}^2(\mathbf{x}_n)} p(\mathbf{x}_n|j_n) d\mathbf{x}_n \right. \\
&\quad \left. - \sum_{n=1}^N \left( \int \frac{\pi(\mathbf{x}_n) g(\mathbf{x}_n)}{\varphi_{\mathcal{P}_n}(\mathbf{x}_n)} p(\mathbf{x}_n|j_n) d\mathbf{x}_n \right)^2 \right] P(j_{1:N}).
\end{aligned} \tag{63}$$

#### A. Variance of the estimators of the MIS schemes

In the following, we analyze the variance of the six MIS schemes discussed through this paper under the assumptions described in Theorem 1 (see Section VI for more details). Since some schemes arise under more than one sampling/weighting combination (see Table III), here we always use the combination that facilitates the analysis.

**1. [R1] Sampling 1 / Weighting 2:** In this scheme, all the terms of the sum in Eq. (60) are independent, so we can use Eq. (62) for computing the variance of  $\hat{I}$ . Substituting  $\varphi_{j_n}(\mathbf{x}_n) = p(\mathbf{x}_n|j_n) = q_{j_n}(\mathbf{x}_n)$  in 62,

$$\begin{aligned}
\text{Var}(\hat{I}_{R1}) &= \frac{1}{Z^2 N^2} \sum_{n=1}^N \sum_{j_n=1}^N \left[ \int \frac{\pi^2(\mathbf{x}_n) g^2(\mathbf{x}_n)}{p^2(\mathbf{x}_n|j_n)} p(\mathbf{x}_n|j_n) P(j_n) d\mathbf{x}_n \right] \\
&\quad - \frac{I^2}{N} \\
&= \frac{1}{Z^2 N^2} \sum_{n=1}^N \left[ \int \sum_{j_n=1}^N \frac{\pi^2(\mathbf{x}_n) g^2(\mathbf{x}_n)}{q_{j_n}(\mathbf{x}_n)} P(j_n) d\mathbf{x}_n \right] - \frac{I^2}{N} \\
&= \frac{1}{Z^2 N^2} \sum_{n=1}^N \left[ \int \frac{1}{N} \sum_{k=1}^N \frac{\pi^2(\mathbf{x}_n) g^2(\mathbf{x}_n)}{q_k(\mathbf{x}_n)} d\mathbf{x}_n \right] - \frac{I^2}{N} \\
&= \frac{1}{Z^2 N^2} \sum_{k=1}^N \int \frac{\pi^2(\mathbf{x}) g^2(\mathbf{x})}{q_k(\mathbf{x})} d\mathbf{x} - \frac{I^2}{N},
\end{aligned} \tag{64}$$

where we have used that  $P(j_n) = \frac{1}{N}$ ,  $\forall j_n \in \{1, \dots, N\}$ .

**2. [R2] Sampling 1 / Weighting 4:** The expression for the conditional independence of Eq. (63) is now used substituting

$\varphi_{j_{1:N}}(\mathbf{x}_n) = f(\mathbf{x}_n|j_{1:N}) = \frac{1}{N} \sum_{k=1}^N q_{j_k}(\mathbf{x}_n)$  and averaging it over the  $N^N$  equiprobable sequences of indexes  $j_{1:N}$ :

$$\begin{aligned}
\text{Var}(\hat{I}_{R2}) &= \frac{1}{Z^2 N^2} \left[ \sum_{j_{1:N}} \left[ \sum_{n=1}^N \int \frac{\pi^2(\mathbf{x}_n) g^2(\mathbf{x}_n)}{\varphi_{j_{1:N}}^2(\mathbf{x}_n)} p(\mathbf{x}_n|j_n) d\mathbf{x}_n \right. \right. \\
&\quad \left. \left. - \sum_{n=1}^N \left( \int \frac{\pi(\mathbf{x}_n) g(\mathbf{x}_n)}{\varphi_{j_{1:N}}(\mathbf{x}_n)} p(\mathbf{x}_n|j_n) d\mathbf{x}_n \right)^2 \right] P(j_{1:N}) \right] \\
&= \frac{1}{Z^2 N^2} \frac{1}{N^N} \left[ \sum_{j_{1:N}} \sum_{n=1}^N \int \frac{\pi^2(\mathbf{x}_n) g^2(\mathbf{x}_n)}{f^2(\mathbf{x}_n|j_{1:N})} q_{j_n}(\mathbf{x}_n) d\mathbf{x}_n \right. \\
&\quad \left. - \sum_{j_{1:N}} \sum_{n=1}^N \left( \int \frac{\pi(\mathbf{x}_n) g(\mathbf{x}_n)}{f(\mathbf{x}_n|j_{1:N})} q_{j_n}(\mathbf{x}_n) d\mathbf{x}_n \right)^2 \right] \\
&= \frac{1}{Z^2 N} \frac{1}{N^N} \left[ \sum_{j_{1:N}} \int \frac{\pi^2(\mathbf{x}) g^2(\mathbf{x})}{f^2(\mathbf{x}|j_{1:N})} \left( \frac{1}{N} \sum_{n=1}^N q_{j_n}(\mathbf{x}) \right) d\mathbf{x} \right. \\
&\quad \left. - \frac{1}{N} \sum_{j_{1:N}} \sum_{n=1}^N \left( \int \frac{\pi(\mathbf{x}_n) g(\mathbf{x}_n)}{f(\mathbf{x}_n|j_{1:N})} q_{j_n}(\mathbf{x}_n) d\mathbf{x}_n \right)^2 \right] \\
&= \frac{1}{Z^2 N} \frac{1}{N^N} \left[ \sum_{j_{1:N}} \int \frac{\pi^2(\mathbf{x}) g^2(\mathbf{x})}{f(\mathbf{x}|j_{1:N})} d\mathbf{x} \right. \\
&\quad \left. - \frac{1}{N} \sum_{j_{1:N}} \sum_{n=1}^N \left( \int \frac{\pi(\mathbf{x}_n) g(\mathbf{x}_n)}{f(\mathbf{x}_n|j_{1:N})} q_{j_n}(\mathbf{x}_n) d\mathbf{x}_n \right)^2 \right].
\end{aligned} \tag{65}$$

where we have used the identity  $f(\mathbf{x}|j_{1:N}) = \frac{1}{N} \sum_{n=1}^N q_{j_n}(\mathbf{x}_n)$ . This expression for the variance resembles that of scheme [N3], averaged over the  $N^N$  possible mixtures (combinations) that can arise with sampling  $\mathcal{S}_1$ .

**3. [R3] Sampling 1 / Weighting 3:** All the elements are independent in the sum, and the weights do not depend on any index of the set  $j_{1:N}$ . Therefore, we can start with Eq. (62), marginalize over the indexes, and substitute  $\varphi_n(\mathbf{x}_n) = p(\mathbf{x}_n) = \psi(\mathbf{x}_n)$ ,

$$\begin{aligned}
\text{Var}(\hat{I}_{R3}) &= \frac{1}{Z^2 N^2} \sum_{n=1}^N \int \frac{\pi^2(\mathbf{x}_n) g^2(\mathbf{x}_n)}{\varphi_n^2(\mathbf{x}_n)} p(\mathbf{x}_n) d\mathbf{x}_n \\
&\quad - \frac{1}{Z^2 N^2} \sum_{n=1}^N \left( \int \frac{\pi(\mathbf{x}_n) g(\mathbf{x}_n)}{\varphi_n(\mathbf{x}_n)} p(\mathbf{x}_n) d\mathbf{x}_n \right)^2 \\
&= \frac{1}{Z^2 N^2} \sum_{n=1}^N \int \frac{\pi^2(\mathbf{x}_n) g^2(\mathbf{x}_n)}{\psi^2(\mathbf{x}_n)} \psi(\mathbf{x}_n) d\mathbf{x}_n \\
&\quad - \frac{1}{Z^2 N^2} \sum_{n=1}^N \left( \int \frac{\pi(\mathbf{x}_n) g(\mathbf{x}_n)}{\psi(\mathbf{x}_n)} \psi(\mathbf{x}_n) d\mathbf{x}_n \right)^2 \\
&= \frac{1}{Z^2 N} \int \frac{\pi^2(\mathbf{x}) g^2(\mathbf{x})}{\psi(\mathbf{x})} d\mathbf{x} - \frac{I^2}{N}.
\end{aligned} \tag{66}$$

**4. [N1] Sampling 3 / Weighting 3:** The methods that use sampling without replacement introduce correlation at the selection of the proposals. However, under the perspective of the deterministic sampling ( $\mathcal{S}_3$ ), the  $n$ -th sample  $\mathbf{x}_n$  is a realization of the r.v.  $X_n \sim q_n$  and is independent of the other samples (see Fig. 6). Marginalizing first Eq. (62) over the indexes, and

substituting  $\varphi_n(\mathbf{x}_n) = p(\mathbf{x}_n) = q_n(\mathbf{x}_n)$ :

$$\begin{aligned}
\text{Var}(\hat{I}_{N1}) &= \frac{1}{Z^2 N^2} \sum_{n=1}^N \int \frac{\pi^2(\mathbf{x}_n) g^2(\mathbf{x}_n)}{\varphi_n^2(\mathbf{x}_n)} p(\mathbf{x}_n) d\mathbf{x}_n \\
&\quad - \frac{1}{Z^2 N^2} \sum_{n=1}^N \left( \int \frac{\pi(\mathbf{x}_n) g(\mathbf{x}_n)}{\varphi_n(\mathbf{x}_n)} p(\mathbf{x}_n) d\mathbf{x}_n \right)^2 \\
&= \frac{1}{Z^2 N^2} \sum_{n=1}^N \int \frac{\pi^2(\mathbf{x}_n) g^2(\mathbf{x}_n)}{q_n^2(\mathbf{x}_n)} q_n(\mathbf{x}_n) d\mathbf{x}_n \\
&\quad - \frac{1}{Z^2 N^2} \sum_{n=1}^N \left( \int \frac{\pi(\mathbf{x}_n) g(\mathbf{x}_n)}{q_n(\mathbf{x}_n)} q_n(\mathbf{x}_n) d\mathbf{x}_n \right)^2 \\
&= \frac{1}{Z^2 N^2} \sum_{n=1}^N \int \frac{\pi^2(\mathbf{x}_n) g^2(\mathbf{x}_n)}{q_n(\mathbf{x}_n)} d\mathbf{x}_n - \frac{I^2}{N}.
\end{aligned} \tag{67}$$

**5. [N2] Sampling 2 / Weighting 1:** In this scheme, we use again the expression for conditional independence of Eq. (63).

Substituting  $\varphi_{j_{1:n-1}} = p(\mathbf{x}_n | j_{1:n-1})$ ,

$$\begin{aligned}
\text{Var}(\hat{I}_{N2}) &= \frac{1}{Z^2 N^2} \sum_{j_{1:N}} \left[ \sum_{n=1}^N \int \frac{\pi^2(\mathbf{x}_n) g^2(\mathbf{x}_n)}{\varphi_{j_{1:n-1}}^2(\mathbf{x}_n)} p(\mathbf{x}_n | j_n) d\mathbf{x}_n \right. \\
&\quad \left. - \sum_{n=1}^N \left( \int \frac{\pi(\mathbf{x}_n) g(\mathbf{x}_n)}{\varphi_{j_{1:n-1}}(\mathbf{x}_n)} p(\mathbf{x}_n | j_n) d\mathbf{x}_n \right)^2 \right] P(j_{1:N}) \\
&= \frac{1}{Z^2 N^2} \sum_{n=1}^N \sum_{j_{1:n}} \left[ \int \frac{\pi^2(\mathbf{x}_n) g^2(\mathbf{x}_n)}{p^2(\mathbf{x}_n | j_{1:n-1})} p(\mathbf{x}_n | j_n) d\mathbf{x}_n \right. \\
&\quad \left. - \left( \int \frac{\pi(\mathbf{x}_n) g(\mathbf{x}_n)}{p(\mathbf{x}_n | j_{1:n-1})} p(\mathbf{x}_n | j_n) d\mathbf{x}_n \right)^2 \right] P(j_{1:n}) \\
&= \frac{1}{Z^2 N^2} \sum_{n=1}^N \sum_{j_{1:n-1}} \int \frac{\pi^2(\mathbf{x}_n) g^2(\mathbf{x}_n)}{p(\mathbf{x}_n | j_{1:n-1})} P(j_{1:n-1}) d\mathbf{x}_n \\
&\quad - \frac{1}{Z^2 N^2} \sum_{n=1}^N \sum_{j_{1:n}} \left( \int \frac{\pi(\mathbf{x}_n) g(\mathbf{x}_n)}{p(\mathbf{x}_n | j_{1:n-1})} q_{j_n} d\mathbf{x}_n \right)^2 P(j_{1:n})
\end{aligned} \tag{68}$$

Since the the integrals only depend on the set of indexes  $j_{1:n}$ , each term of the sum has been first marginalized over  $j_{n+1:N}$ . The first term in the sum can then be further marginalized over  $j_n$  to obtain the final expression. Note that the variance is the average of the variance of all the  $N!$  possible sequences of indexes in the sampling without replacement.

**6. [N3] Sampling 3 / Weighting 5:** We have followed the same arguments of scheme N1. Marginalizing Eq. (62) over all

the set of indexes  $j_{1:N}$ , and substituting  $\varphi_n(\mathbf{x}_n) = f(\mathbf{x}_n) = \psi(\mathbf{x}_n)$ :

$$\begin{aligned}
\text{Var}(\hat{I}_{N3}) &= \frac{1}{Z^2 N^2} \sum_{n=1}^N \int \frac{\pi^2(\mathbf{x}_n) g^2(\mathbf{x}_n)}{\psi^2(\mathbf{x}_n)} q_n(\mathbf{x}_n) d\mathbf{x}_n \\
&\quad - \frac{1}{Z^2 N^2} \sum_{n=1}^N \left( \int \frac{\pi(\mathbf{x}_n) g(\mathbf{x}_n)}{\psi(\mathbf{x}_n)} q_n(\mathbf{x}_n) d\mathbf{x}_n \right)^2 \\
&= \frac{1}{Z^2 N} \int \frac{\pi^2(\mathbf{x}) g^2(\mathbf{x})}{\psi^2(\mathbf{x})} \left( \frac{1}{N} \sum_{n=1}^N q_n(\mathbf{x}) \right) d\mathbf{x} \\
&\quad - \frac{1}{Z^2 N^2} \sum_{n=1}^N \left( \int \frac{\pi(\mathbf{x}) g(\mathbf{x})}{\psi(\mathbf{x})} q_n(\mathbf{x}) d\mathbf{x} \right)^2 \\
&= \frac{1}{Z^2 N} \int \frac{\pi^2(\mathbf{x}) g^2(\mathbf{x})}{\psi(\mathbf{x})} d\mathbf{x} \\
&\quad - \frac{1}{Z^2 N^2} \sum_{n=1}^N \left( \int \frac{\pi(\mathbf{x}) g(\mathbf{x})}{\psi(\mathbf{x})} q_n(\mathbf{x}) d\mathbf{x} \right)^2,
\end{aligned} \tag{69}$$

where we have used the identity  $\psi(\mathbf{x}) = \frac{1}{N} \sum_{n=1}^N q_n(\mathbf{x}) d\mathbf{x}$ .

### B. Proof of Theorem 1

The proof of Theorem 1 is split in the next three propositions.

**Proposition 1.**  $\text{Var}(\hat{I}_{R1}) = \text{Var}(\hat{I}_{N1})$

**Proof:** See that Eqs. (64) and (67) are equivalent. □

**Proposition 2.**  $\text{Var}(\hat{I}_{N1}) \geq \text{Var}(\hat{I}_{R3})$ .

**Proof:** Subtracting Eqs. (66) and (67), we get

$$\begin{aligned}
\text{Var}(\hat{I}_{R3}) - \text{Var}(\hat{I}_{N1}) &= \\
&= \frac{1}{Z^2 N^2} \int \left( \frac{N}{\frac{1}{N} \sum_{j=1}^N q_j(\mathbf{x})} - \sum_{i=1}^N \frac{1}{q_i(\mathbf{x})} \right) g^2(\mathbf{x}) \pi^2(\mathbf{x}) d\mathbf{x}.
\end{aligned}$$

Since  $g^2(\mathbf{x}) \pi^2(\mathbf{x}) \geq 0 \forall \mathbf{x} \in \mathbb{R}^{d_x}$ , it is sufficient to show that

$$\frac{1}{\frac{1}{N} \sum_{j=1}^N q_j(\mathbf{x})} \leq \frac{1}{N} \sum_{i=1}^N \frac{1}{q_i(\mathbf{x})}. \tag{70}$$

Now, let us note that the left-hand side of Eq. (70) is the inverse of the arithmetic mean of  $q_1(\mathbf{x}), \dots, q_N(\mathbf{x})$ ,

$$A_N = \frac{1}{N} \sum_{j=1}^N q_j(\mathbf{x}),$$

whereas the right hand side of Eq. (70) is the inverse of the harmonic mean of  $q_1(\mathbf{x}), \dots, q_N(\mathbf{x})$ ,

$$\frac{1}{H_N} = \frac{1}{N} \sum_{i=1}^N \frac{1}{q_i(\mathbf{x})}.$$

Therefore, the inequality in Eq. (70) is equivalent to stating that  $\frac{1}{A_N} \leq \frac{1}{H_N}$ , or equivalently  $A_N \geq H_N$ , which is the well-known arithmetic mean–harmonic mean inequality for positive real numbers [24, 25, 26]. □

**Proposition 3.**  $\text{Var}(\hat{I}_{R3}) \geq \text{Var}(\hat{I}_{N3})$ .

**Proof:** Subtracting (66) and (69), we get

$$\text{Var}(\hat{I}_{N3}) - \text{Var}(\hat{I}_{R3}) = -\frac{I^2}{N} + \frac{1}{Z^2 N^2} \sum_{n=1}^N \left( \int \frac{\pi(\mathbf{x})g(\mathbf{x})}{\psi(\mathbf{x})} q_n(\mathbf{x}) d\mathbf{x} \right)^2$$

Therefore, the proposition is proved if

$$\frac{1}{Z^2} \sum_{n=1}^N \left( \int \frac{\pi(\mathbf{x})g(\mathbf{x})}{\psi(\mathbf{x})} q_n(\mathbf{x}) d\mathbf{x} \right)^2 \geq N I^2$$

If we substitute  $I$  with the expression of Eq. (27), multiplying both numerator and denominator by  $\psi(\mathbf{x})$  in the integral of the right-hand side,

$$\begin{aligned} \frac{1}{Z^2} \sum_{n=1}^N \left( \int \frac{\pi(\mathbf{x})g(\mathbf{x})}{\psi(\mathbf{x})} q_n(\mathbf{x}) d\mathbf{x} \right)^2 &\geq N \left( \frac{1}{Z} \int \frac{\pi(\mathbf{x})g(\mathbf{x})}{\psi(\mathbf{x})} \psi(\mathbf{x}) d\mathbf{x} \right)^2 \\ \sum_{n=1}^N \left( \int \frac{\pi(\mathbf{x})g(\mathbf{x})}{\psi(\mathbf{x})} q_n(\mathbf{x}) d\mathbf{x} \right)^2 &\geq N \left( \int \frac{\pi(\mathbf{x})g(\mathbf{x})}{\psi(\mathbf{x})} \left( \frac{1}{N} \sum_{n=1}^N q_n(\mathbf{x}) \right) d\mathbf{x} \right)^2 \\ \sum_{n=1}^N \left( \int \frac{\pi(\mathbf{x})g(\mathbf{x})}{\psi(\mathbf{x})} q_n(\mathbf{x}) d\mathbf{x} \right)^2 &\geq \frac{1}{N} \left( \sum_{n=1}^N \int \frac{\pi(\mathbf{x})g(\mathbf{x})}{\psi(\mathbf{x})} q_n(\mathbf{x}) d\mathbf{x} \right)^2 \\ N \sum_{n=1}^N a_n^2 &\geq \left( \sum_{n=1}^N a_n \right)^2 \end{aligned} \quad (71)$$

with  $a_n = \int \frac{\pi(\mathbf{x})g(\mathbf{x})}{\psi(\mathbf{x})} q_n(\mathbf{x}) d\mathbf{x}$ . The inequality of Eq. (71) holds, since it is the definition of the Cauchy-Schwarz inequality [24],

$$\left( \sum_{n=1}^N a_n^2 \right) \left( \sum_{n=1}^N b_n^2 \right) \geq \left( \sum_{n=1}^N a_n b_n \right)^2, \quad (72)$$

with  $b_n = 1$  for  $n = 1, \dots, N$ . □

**Proof of Theorem 1.** The proof is obtained by applying Propositions 1, 2, and 3. □

### C. Proof of Theorem 2

Let us first particularize the variance expression for  $N = 2$ . From Eq. (67),

$$\begin{aligned} \text{Var}(\hat{I}_{N1}) &= \text{Var}(\hat{I}_{R1}) \\ &= \frac{1}{4Z^2} \left( \int \frac{\pi^2(\mathbf{x})g^2(\mathbf{x})}{q_1(\mathbf{x})} d\mathbf{x} + \int \frac{\pi^2(\mathbf{x})g^2(\mathbf{x})}{q_2(\mathbf{x})} d\mathbf{x} \right) - \frac{I^2}{2}. \end{aligned} \quad (73)$$

From Eq. (66),

$$\text{Var}(\hat{I}_{R3}) = \frac{1}{2Z^2} \int \frac{\pi^2(\mathbf{x})g^2(\mathbf{x})}{\frac{q_1(\mathbf{x})+q_2(\mathbf{x})}{2}} d\mathbf{x} - \frac{I^2}{2}. \quad (74)$$

From Eq. (69),

$$\begin{aligned} \text{Var}(\hat{I}_{N3}) &= \frac{1}{2Z^2} \int \frac{\pi^2(\mathbf{x})g^2(\mathbf{x})}{\frac{q_1(\mathbf{x})+q_2(\mathbf{x})}{2}} d\mathbf{x} \\ &\quad - \frac{1}{4Z^2} \left( \int \frac{\pi(\mathbf{x})g(\mathbf{x})}{\frac{q_1(\mathbf{x})+q_2(\mathbf{x})}{2}} q_1(\mathbf{x}) d\mathbf{x} \right)^2 \\ &\quad - \frac{1}{4Z^2} \left( \int \frac{\pi(\mathbf{x})g(\mathbf{x})}{\frac{q_1(\mathbf{x})+q_2(\mathbf{x})}{2}} q_2(\mathbf{x}) d\mathbf{x} \right)^2. \end{aligned} \quad (75)$$

From Eq. (65),

$$\begin{aligned}
\text{Var}(\hat{I}_{R2}) &= \frac{1}{8Z^2} \left( \int \frac{\pi^2(\mathbf{x})g^2(\mathbf{x})}{q_1(\mathbf{x})} d\mathbf{x} + \int \frac{\pi^2(\mathbf{x})g^2(\mathbf{x})}{q_2(\mathbf{x})} d\mathbf{x} \right) - \frac{1}{4}I^2 \\
&\quad + \frac{1}{4Z^2} \int \frac{\pi^2(\mathbf{x})g^2(\mathbf{x})}{\frac{q_1(\mathbf{x})+q_2(\mathbf{x})}{2}} d\mathbf{x} \\
&\quad - \frac{1}{8Z^2} \left( \int \frac{\pi(\mathbf{x})g(\mathbf{x})}{\frac{q_1(\mathbf{x})+q_2(\mathbf{x})}{2}} q_1(\mathbf{x}) d\mathbf{x} \right)^2 \\
&\quad - \frac{1}{8Z^2} \left( \int \frac{\pi(\mathbf{x})g(\mathbf{x})}{\frac{q_1(\mathbf{x})+q_2(\mathbf{x})}{2}} q_2(\mathbf{x}) d\mathbf{x} \right)^2.
\end{aligned} \tag{76}$$

From Eq. (68),

$$\begin{aligned}
\text{Var}(\hat{I}_{N2}) &= \frac{1}{4Z^2} \int \frac{\pi^2(\mathbf{x})g^2(\mathbf{x})}{\frac{q_1(\mathbf{x})+q_2(\mathbf{x})}{2}} d\mathbf{x} + \frac{1}{8Z^2} \int \frac{\pi^2(\mathbf{x})g^2(\mathbf{x})}{q_1(\mathbf{x})} d\mathbf{x} \\
&\quad + \frac{1}{8Z^2} \int \frac{\pi^2(\mathbf{x})g^2(\mathbf{x})}{q_2(\mathbf{x})} d\mathbf{x} \\
&\quad - \frac{1}{8Z^2} \left( \int \frac{\pi(\mathbf{x})g(\mathbf{x})}{\frac{q_1(\mathbf{x})+q_2(\mathbf{x})}{2}} q_1(\mathbf{x}) d\mathbf{x} \right)^2 \\
&\quad - \frac{1}{8Z^2} \left( \int \frac{\pi(\mathbf{x})g(\mathbf{x})}{\frac{q_1(\mathbf{x})+q_2(\mathbf{x})}{2}} q_2(\mathbf{x}) d\mathbf{x} \right)^2 - \frac{I^2}{4}.
\end{aligned} \tag{77}$$

**Proposition 4.** For  $N = 2$ ,  $\text{Var}(\hat{I}_{R2}) = \text{Var}(\hat{I}_{N2})$

**Proof:** See that Eqs. (76) and (77) are equivalent.  $\square$

**Proposition 5.** For  $N = 2$ ,  $\text{Var}(\hat{I}_{N1}) \geq \text{Var}(\hat{I}_{R2}) \geq \text{Var}(\hat{I}_{N3})$

**Proof:** Analyzing Eqs. (73) and (75), we see that Eq. (76) can be rewritten as

$$\text{Var}(\hat{I}_{R2}) = \frac{1}{2} \text{Var}(\hat{I}_{N1}) + \frac{1}{2} \text{Var}(\hat{I}_{N3}). \tag{78}$$

Since in Theorem 1 it is proved that  $\text{Var}(\hat{I}_{N1}) \geq \text{Var}(\hat{I}_{N3})$  for any  $N$ , the proposition holds at least for  $N = 2$ .  $\square$

**Proof of Theorem 2.** The proof is obtained by applying Propositions 4 and 5.  $\square$

**Remark 3.** We hypothesize that Theorem 2 might also hold for  $N > 2$ . The MIS schemes R2 and N2 seem to average estimators with variance reduction (related to N3) with estimators with worse variance (related to N1).

**Remark 4.** Note that the scheme R3 does not appear in Theorem 2. Eq. (76) can be rewritten as

$$\begin{aligned}
\text{Var}(\hat{I}_{R2}) &= \frac{1}{2} \text{Var}(\hat{I}_{R3}) + \\
&\quad \frac{1}{8Z^2} \left( \int \frac{\pi^2(\mathbf{x})g^2(\mathbf{x})}{q_1(\mathbf{x})} d\mathbf{x} + \int \frac{\pi^2(\mathbf{x})g^2(\mathbf{x})}{q_2(\mathbf{x})} d\mathbf{x} \right) \\
&\quad - \frac{1}{8Z^2} \left( \int \frac{\pi(\mathbf{x})g(\mathbf{x})}{\frac{q_1(\mathbf{x})+q_2(\mathbf{x})}{2}} q_1(\mathbf{x}) d\mathbf{x} \right)^2 \\
&\quad - \frac{1}{8Z^2} \left( \int \frac{\pi(\mathbf{x})g(\mathbf{x})}{\frac{q_1(\mathbf{x})+q_2(\mathbf{x})}{2}} q_2(\mathbf{x}) d\mathbf{x} \right)^2.
\end{aligned}$$

The question is then whether the last four terms are larger than  $\frac{1}{2} \text{Var}(\hat{I}_{R3})$ . We hypothesize that no inequality can be established in a general case, i.e., whether the scheme R3 would outperform R2 or not for a given  $\pi(\mathbf{x})$  and  $g(\mathbf{x})$ , might depend on the proposals  $q_1(\mathbf{x})$  and  $q_2(\mathbf{x})$ .